Silvano Salvador

Toponomastic Entropy

An Information-Theoretic and Bayesian Framework for the Detection of Linguistic Substrates in Toponomastic Corpora

Volume I — Methodological Foundations

Complete Theoretical and Computational Treatment

Contents

Preface **vi**

1 Toponyms as Phonotactic Fossils **1**
1.1 The Persistence of Toponymic Strata . . . 1
1.2 Why Information Theory? . . . 2
1.3 Historical Background: Quantitative Approaches to Toponymy . . . 3
1.3.1 Early Statistical Approaches . . . 3
1.3.2 Computational Linguistics and NLP Approaches . . . 3
1.3.3 The German-Language Toponymic Tradition . . . 4
1.3.4 Information Theory in Linguistics . . . 4
1.4 Typological and Geographic Scope . . . 5
1.4.1 Iberia: The Benchmark Region . . . 5
1.4.2 Central Europe: Extinct and Partially Attested Substrates . . . 5
1.4.3 British Isles: Overlapping Celtic Varieties . . . 5
1.4.4 Indian Subcontinent: The Largest Corpus . . . 6
1.4.5 Anatolia and the Caucasus: Maximum Stratification . . . 6
1.5 Overview of the Series . . . 6

2 Mathematical Foundations **8**
2.1 Probability Distributions over n-Gram Space . . . 8
2.2 Shannon Entropy of a Toponymic Corpus . . . 9
2.3 Conditional Entropy and the Chain Rule . . . 10
2.4 Perplexity and Its Relation to Cross-Entropy . . . 12
2.5 Kullback–Leibler Divergence and Related Measures . . . 13
2.6 Jensen–Shannon Divergence as a Metric . . . 14
2.7 Rényi Entropy and Its Role as a Generalisation . . . 16
2.8 The Mixing Entropy Lemma . . . 19
2.9 Mutual Information and Spatial Gradients . . . 20

3 Statistical Estimation and Hypothesis Testing **22**
3.1 The Multinomial Sampling Model . . . 22
3.2 Confidence Intervals for Entropy Estimates . . . 22
3.2.1 Derivation of the Miller–Madow Correction . . . 23
3.2.2 The NSB (Nemenman–Shafee–Bialek) Estimator . . . 24
3.2.3 Power Analysis for Entropy Contrasts . . . 25
3.3 Permutation Tests for KL Divergence . . . 25
3.4 Simulation Studies . . . 26
3.5 Multiple Comparison Correction for Pairwise Zone Tests . . . 27

3.6 Bootstrap Confidence Intervals for KL Divergence 28
3.7 Treatment of Imbalanced Zone Corpora 29
3.8 Spatial Dependence and the Validity of Permutation Tests 31

4 Computational Pipeline **34**
4.1 Data Acquisition: The GeoNames Database 34
4.1.1 Feature Class Filtering 34
4.1.2 Alternate Names and Coordinate Completeness 34
4.1.3 Toponym Normalisation 36
4.1.4 Generic Stoplist Design 37
4.2 Feature Extraction 38
4.3 Complete Pipeline: Pseudocode Specification 38
4.4 TF-IDF Weighting 38
4.5 Dimensionality Reduction and Visualisation 40
4.5.1 Principal Component Analysis 40
4.5.2 UMAP Embedding 40
4.6 Unsupervised Clustering 40
4.6.1 Alternative Clustering Methods 41
4.7 Bayesian Model Comparison 42
4.7.1 Dirichlet-Multinomial Marginal Likelihood 42
4.7.2 Bayesian Isolation Score 42

5 Validation on the Iberian Prototype **46**
5.1 Corpus Description and Design 46
5.2 Phonotactic Entropy Results 47
5.3 KL Divergence Matrix 47
5.4 Unsupervised Clustering and Spatial Validation 49
5.4.1 Interpreting the Purity Figure 49
5.4.2 Spatial Coherence of Cluster Assignments 49
5.4.3 Bayesian Isolation Score per Zone 50
5.4.4 Perplexity Cross-Validation 51
5.4.5 k-Fold Cross-Validation of the Zone Model 51
5.4.6 Sensitivity of Entropy Estimates to Zone Boundary Definitions 53

6 Methodological Extensions **56**
6.1 Optimal n-gram Order Selection 56
6.2 Higher-Order Markov Models 56
6.3 Hierarchical Substrate Detection 57
6.4 Temporal Stratification via Document Dating 57
6.5 Spatial Substrate Mapping 58
6.6 Network-Based Substrate Modelling 58
6.6.1 Network Centrality as a Substrate Diagnostic 58
6.6.2 Spectral Properties of the Transition Matrix 59
6.7 Semi-Supervised Substrate Assignment 60
6.8 Robustness to Script and Transliteration 61
6.9 Relation to Traditional Substrate Detection Methods 61
6.9.1 Etymological Substrate Analysis 62
6.9.2 Lexical Density Mapping 63

6.9.3 Pellegrini's Linguistic-Stratigraphy Model 63
6.10 Smoothing Estimators: A Theoretical Comparison 65
6.11 Information-Theoretic Bounds on Classifier Accuracy 67
6.12 Optimal Transport and Wasserstein Distances Between Phonotactic Distributions 68

7 Spatial Autocorrelation and Geographic Mutual Information 72
7.1 Spatial Autocorrelation of Phonotactic Entropy 72
7.1.1 Statistical Power of the Global Test 73
7.2 Local Indicators of Spatial Association (LISA) 74
7.3 Spatial Mutual Information 74
7.3.1 Variogram Analysis of Pairwise KL Divergence 75
7.4 Implications for Substrate Boundary Detection 76
7.5 Choice of Spatial Weights Matrix 76
7.6 Edge Effects and the Modifiable Areal Unit Problem 78

8 Systematic Morphological Comparison of the Six Onomastic Systems 80
8.1 A Typology of Toponymic Morphology 80
8.2 Iberian Zones: Fusional Romance vs. Agglutinative Basque 81
8.3 British Isles: Celtic Mutation Systems and Romance-Anglo-Saxon Layering 81
8.4 Anatolia–Caucasus: Agglutinative Turkic vs. Fusional Indo-European vs. Ergative Caucasian 82
8.5 Cross-System Synthesis: Morphological Type as a Predictor of Entropy Rank 83
8.6 Indian Subcontinent: Indo-Aryan Compounding, Dravidian Suffixation, and Munda Roots 83
8.7 Central Europe: Extinct Substrates and the Depth of Romanisation . 85
8.8 Etymology-Based Substrate Assignment 86
8.9 Typological Distance Measures 87
8.10 Automated Language Identification 87

9 Sensitivity Analysis and Robustness 88
9.1 Sensitivity to Corpus Size 88
9.2 Sensitivity to Normalisation Parameters 88
9.2.1 Compound Sensitivity: Combined Variations 89
9.2.2 Sensitivity to Vocabulary Cutoff 90
9.3 Jackknife Variance Estimation 90
9.4 Cross-Validation via Leave-One-Zone-Out 91
9.4.1 Full Leave-One-Zone-Out Results 91
9.5 Non-Parametric Permutation Bounds on Estimation Error 92
9.6 Asymptotic Distribution of the Plug-in KL Estimator 93
9.7 Sensitivity to the Choice of Divergence Measure 94

10 Conclusions and Series Roadmap 96
10.1 Summary of Contributions 96
10.2 What the Framework Detects and What It Does Not 96
10.3 Series Architecture 97

10.3.1 Calibration Hierarchy Across Volumes . . . 98
10.4 Open Problems and Future Directions . . . 98
10.4.1 Volume VII: Maghreb . . . 99
10.4.2 Supervised and Active-Learning Extensions . . . 100
10.4.3 Connection to Dialectometric and Language-Distance Research 100
10.5 Reflections on Methodology: Quantification and Interpretation . . . 101
10.6 Final Remarks . . . 102

11 Cross-Regional Comparative Analysis 104
11.1 Master Summary Table . . . 104
11.2 Between-Region Patterns . . . 105
11.2.1 Entropy Range by Region . . . 105
11.2.2 Universal Cross-Regional Entropy Bounds . . . 105
11.2.3 Clustering Purity: A Cross-Volume Comparison . . . 106
11.2.4 Bayes Factors and the Decisive Evidence Threshold . . . 106
11.2.5 The Substrate Isolation Criterion Across Regions . . . 106
11.2.6 Phonotactic Entropy and Language Family: A Tentative Typology . . . 107
11.3 Methodological Conclusions from the Cross-Regional Survey . . . 107
11.4 Cross-Volume Perplexity Analysis . . . 108
11.5 Conditional Entropy and n-gram Predictability . . . 109
11.6 Trigram vs. Bigram Divergence: Additional Discriminative Power . . . 110
11.7 Conditional Entropy as a Predictor of Cross-Volume Perplexity Behaviour . . . 110
11.8 Entropy Bar Charts . . . 111
11.9 KL Divergence Heatmaps . . . 111
11.10 PCA Projection Plots . . . 111
11.11 Geographic Distribution Maps . . . 112
11.12 Ward Dendrogram: Cross-Volume Hierarchical Clustering . . . 112

A Proof Compendium 131
A.1 Properties of Shannon Entropy . . . 131
A.2 KL Divergence: Non-negativity and Convexity . . . 132
A.3 Jensen–Shannon Divergence . . . 133
A.4 Entropy Estimation: Bias and Consistency . . . 134
A.5 Consistency of the Plug-in KL Estimator . . . 135
A.6 Bootstrap Confidence Intervals for Entropy . . . 136
A.7 Permutation Test for KL Distinctiveness . . . 136
A.8 Notation Index . . . 136

B GeoNames Corpus Summary by Volume 138
B.1 Notes on Country-Level Data Quality . . . 138

C Supplementary Statistical Tables 140
C.1 Conditional Entropy by Volume . . . 140
C.2 Perplexity Matrix: Anatolia–Caucasus . . . 141

Marker Bigram and Top-Frequency Bigram Tables **142**
C.3 Marker Bigram Frequencies: British Isles (Vol. IV) 142
C.4 Top-30 Bigrams by Zone: Anatolia and Caucasus (Vol. VI) 142
C.5 Top-10 Bigrams by Zone: Maghreb (Vol. VII) 144

Asymptotic Theory of the Permutation Framework **146**
C.6 Setup and Notation 146
C.7 Large-Sample Distribution Under the Null 146
C.7.1 Asymptotic Chi-Squared Limit 146
C.8 Rate of Convergence of the Permutation Distribution 147
C.8.1 Berry–Esseen Bound 147
C.9 Multiple Testing Correction in the Context of Permutation Tests . . . 148
C.10 Power Under Sparse Alternatives 148
C.10.1 Formulation of the Sparse Alternative 148
C.11 Jackknife Variance Estimation for the KL Divergence 149
C.11.1 Definition and Computational Cost 149
C.11.2 Comparison with Bootstrap Variance at Finite N 150
C.12 Summary and Practical Recommendations 150

Preface

The present volume inaugurates a series devoted to a systematic application of information-theoretic and Bayesian methods to one of the oldest problems in historical linguistics: the detection and characterisation of linguistic substrates preserved in toponymic corpora.

Toponyms are, in a precise sense, *fossils of language*. Unlike lexical or phonological evidence, which is continuously eroded by language shift and standardisation, place names exhibit a remarkable conservatism across centuries and even millennia. The name of a river, a mountain pass, or a hamlet may transmit phonotactic patterns belonging to a language that has left no other trace. The *Alt-Europäische Hydronymie* of Hans Krahe, the pre-Roman substrate studies of Pellegrini and Alinei, the Basque substrate research of Mitxelena — all rest ultimately on the intuition that phonotactic patterns persist in place names long after the speech community that coined them has vanished. The present series is an attempt to formalise and quantify that intuition.

The methodological novelty is to treat the phonotactic signature of toponymic corpora as a probability distribution over character n-gram space, and to deploy the full apparatus of information theory — Shannon entropy, Kullback–Leibler divergence, cross-entropy, and mutual information — together with unsupervised clustering and Bayesian model comparison, in order to (i) quantify the internal diversity of a corpus, (ii) measure the phonotactic distance between corpora, and (iii) identify latent structure without presupposing any substrate hypothesis.

This first volume lays the complete mathematical and computational foundations. It is deliberately self-contained: a reader with graduate-level training in probability theory and some familiarity with historical linguistics should be able to follow every argument from definition to proof to computational implementation without consulting secondary sources. Subsequent volumes apply the framework to specific geographic regions: the Iberian Peninsula (Vol. II), Central Europe (Vol. III), the British Isles (Vol. IV), the Indian Subcontinent (Vol. V), and Anatolia–Caucasus (Vol. VI).

Chapter 1

Toponyms as Phonotactic Fossils

1.1 The Persistence of Toponymic Strata

Place names are among the most durable artefacts of human culture. Linguistic communities that have long since disappeared from the archaeological and textual record continue to leave their imprint on the landscape through the place names they assigned to rivers, mountains, settlements, and natural features. This phenomenon — the *substrate persistence* of toponyms — has been recognised since the foundational work of Pellegrini (1990) and, in the pre-Indo-European context, of Krahe (1964).

Substrate persistence operates through a simple mechanism. Once a place name becomes institutionalised — recorded in administrative documents, canonised in oral tradition, reproduced on maps — it acquires a degree of cultural inertia that makes it resistant to the linguistic shifts affecting the spoken language of the community. A Roman legionary stationed in Noricum learned the Celtic name of the local river not as a token of the dying Celtic language but as the name of that particular river; his administrative successors, speaking Latin, then Medieval Latin, then early Rhaeto-Romance, preserved the name through generations of transmission, adapting its phonology to each successive speech variety but never fully erasing the distributional trace of its Celtic origin.

The concept of *toponymic stratigraphy* (Pulgram, 1954; Nicolaisen, 1976) formalises this observation: just as a geological section reveals successive layers of deposition, a toponymic corpus reveals successive layers of linguistic occupation. The deepest layers correspond to the oldest attested or inferable languages; the uppermost to the current dominant tongue. In many parts of Europe, three to five distinct substrate layers can be identified: pre-Indo-European, Celtic or other Indo-European, Latin, medieval vernacular, and modern standard.

Definition: Toponymic Stratigraphy

A *toponymic stratigraphy* for a geographic region $\mathcal{R}$ is a partially ordered sequence of linguistic layers $L_1 \prec L_2 \prec \cdots \prec L_k$ (where L_1 is oldest), such that place names currently attested in $\mathcal{R}$ can be assigned, with varying certainty, to one or more layers based on their phonological, morphological, or etymological properties.

The challenge for quantitative study is that this assignment is, in the traditional approach, based on individual etymological analysis — a procedure that is simul-

taneously labour-intensive, language-dependent, and susceptible to circularity. If a researcher suspects that a particular region harbours a pre-Celtic substrate, they will look for names that do not admit Celtic or Latin explanations and assign them to the hypothesised pre-Celtic layer; the resulting "evidence" for the pre-Celtic layer is then precisely as strong as the researcher's prior belief in it.

1.2 Why Information Theory?

Information theory provides a natural and assumption-free language for describing the internal structure of probability distributions. When we model a toponymic corpus as a distribution over n-gram space, we gain access to a rich set of measures that quantify different aspects of distributional structure.

The fundamental insight motivating the approach is this: if a toponymic corpus contains names from two distinct substrate layers, and those layers differ in their phonological inventory or phonotactic constraints, then the aggregate n-gram distribution of the corpus will be a mixture of two distinct component distributions. The mixture will exhibit higher entropy than either component alone (the *mixing lemma* of information theory), and unsupervised decomposition methods will, under favourable conditions, be able to identify the components without any prior knowledge of their linguistic identity.

(i) **Shannon entropy** H quantifies the internal diversity of a corpus: a corpus with high entropy is phonotactically diverse, while one with low entropy concentrates its probability mass on a small set of recurrent patterns. The entropy of a mixed corpus exceeds the weighted average of its component entropies.

(ii) **Kullback–Leibler divergence** D_{KL} measures the asymmetric distance between two distributions, capturing how surprising one corpus is when modelled by another. If two substrate layers have low mutual KL divergence, they are phonotactically similar and difficult to discriminate; if the divergence is high, they are easy to separate.

(iii) **Cross-entropy** $H_{\times}$ provides a decomposition of the KL divergence into a self-entropy component and a divergence component, useful for understanding the relative contributions of intrinsic diversity and inter-substrate distance.

(iv) **Mutual information** I between positional n-gram distributions and geographic coordinates reveals spatial gradients in phonotactic substrate, distinguishing gradual mixing from sharp areal boundaries.

(v) **Bayesian model comparison** via Bayes factors provides a formally principled framework for evaluating whether a k-substrate model fits the corpus significantly better than a null single-source model, accounting for model complexity through the marginal likelihood.

Remark: On the Absence of Etymology

The framework is deliberately *anetiological*: no etymological identification is assumed or required. The substrate signal is sought in the aggregate distributional

properties of the corpus, not in the analysis of individual names. This is both a methodological strength — it avoids the circularity of etymology-based substrate assignment — and a limitation: the framework identifies *that* a substrate exists and *where* it is strongest, but does not by itself identify *which* language produced it. Identification requires additional evidence, typically from comparative linguistics and archaeology.

1.3 Historical Background: Quantitative Approaches to Toponymy

The application of quantitative methods to toponymic analysis has a history extending back to the mid-twentieth century, though the full potential of information-theoretic approaches has not previously been exploited.

1.3.1 Early Statistical Approaches

The earliest quantitative work on toponymy concentrated on frequency distributions of name-forming elements. Krahe (1964) observed that certain river-name roots (such as **al-*, **ar-*, **isa-*) recurred with striking frequency across a geographically continuous area spanning Western and Central Europe, and used this frequency pattern to argue for a common *Alt-Europäische* stratum predating both Celtic and Germanic settlement. Krahe's argument was essentially statistical — it relied on the improbability of independent parallel coinage of the same root — but it lacked a formal probabilistic framework.

Nicolaisen (1976), in his study of Scottish place names, introduced a more explicit frequency-based methodology, comparing the relative densities of name-forming elements across administrative regions to map the spatial extent of successive linguistic strata (P-Celtic, Q-Celtic, Pictish, Norse, English). His density maps are, in retrospect, a precursor to the spatial mutual information approach developed in Chapter 6 of the present volume.

1.3.2 Computational Linguistics and NLP Approaches

The development of computational linguistics in the 1990s and 2000s brought new tools to the study of toponymy, primarily in the form of automatic named entity recognition and geographic information retrieval. However, the vast majority of this work was concerned with the *identification* of place names in text, not with their *phonological analysis* as carriers of historical substrate information.

The specific application of n-gram language models to toponymic substrate detection appears to have no established precedent in the literature prior to the present series. The closest analogues are found in the related field of *language identification from text*, where character n-gram models are widely used to distinguish between languages (Cavnar and Trenkle, 1994; Dunning, 1994). The present framework inverts this relationship: rather than using a known language model to identify an unknown text, it uses unknown (unlabelled) geographic data to reconstruct the probability distributions characteristic of historical substrate languages.

1.3.3 The German-Language Toponymic Tradition

Independent of and largely parallel to the British and Francophone research traditions, the German-language community developed its own systematic toponymic scholarship during the nineteenth and twentieth centuries, oriented in particular toward the problems of substrate detection in the German-speaking lands and adjacent regions. The tradition is relevant to the present series for two reasons: it is the source of the most influential pre-quantitative substrate hypothesis (Krahe's *Alt-Europäische Hydronymie*), and it developed the most sophisticated manual frequency-analysis tools available prior to the advent of computational methods.

Krahe's (1964) argument for a pan-European pre-Celtic and pre-Germanic hydronymial stratum rests on the observation that a set of approximately sixty river-name roots recurs, with phonological regularity consistent with Proto-Indo-European sound changes, across a geographically continuous zone extending from the British Isles to the Pontic steppe. The argument is explicitly distributional: Krahe does not merely list cognates, but argues from the *density* and *geographic coherence* of their distribution — an argument that, as noted above, anticipates the spatial mutual information analysis of Chapter 7.

The response to Krahe in the German-language literature (the so-called *Krahe-Debatte* of the 1970s–1990s) generated a substantial body of methodologically sophisticated critique, including arguments about the statistical independence of the recurrent roots, the boundary between inheritance and parallel formation, and the representativeness of the hydronymial sample relative to the full toponymic inventory. These methodological critiques translate directly into concerns that the information-theoretic framework addresses formally: independence (handled by the permutation test), parallel formation (equivalent to false positive rate in the etymological classifier), and representativeness (handled by the feature-class filtering and stratified sampling of Chapter 4).

1.3.4 Information Theory in Linguistics

Information theory has been applied to linguistic problems since Shannon's (1951) foundational paper on the entropy of English. Subsequent work has used entropy measures to characterise the typological diversity of the world's languages (Cysouw and Wälchli, 2007), to study lexical complexity (Bentz et al., 2017), and to model language change (Monaghan et al., 2017). The application to historical substrate detection in toponymy is, to the best of the author's knowledge, original to the present series.

Among the information-theoretic tools applied to language at the corpus level, the closest precursor to the present framework is Shannon's prediction experiment, which estimated the entropy of printed English text (approximately 1–1.3 bits per character) by exploiting human prior knowledge of the language to compress test passages. Shannon's experiment established that entropy can be estimated from finite samples of real language, even without access to the full distribution, and that the resulting measure carries interpretable information about the statistical structure of the language. The present series applies this insight to the cross-corpus comparison problem: not the absolute entropy of a single language, but the differential entropy between corpora, used to detect the presence of an admixed second distribu-

tion.

1.4 Typological and Geographic Scope

The choice of the six regions examined in this series was motivated by a combination of substrate complexity, data availability, and methodological diversity. Each region poses a distinct challenge to the framework and is expected, on independent historical-linguistic grounds, to yield a non-trivial substrate signal. A brief overview of the expected substrate configurations, and the specific challenges each region poses, helps situate the methodological choices made in subsequent chapters.

1.4.1 Iberia: The Benchmark Region

The Iberian Peninsula is the natural benchmark for any quantitative substrate methodology. Its substrate configuration is well-established by centuries of scholarship: a pre-Indo-European isolate (Basque) with clear typological distinctiveness from all surrounding languages, an extensive medieval Arabic-influenced zone concentrated in the south and east, and a range of Romance varieties (Castilian, Catalan, Galician-Portuguese) that are phonotactically similar to one another but distinguishable by substrate analysis. The presence of a "positive control" in the form of the Basque zone — whose pre-Indo-European status is beyond scholarly dispute — makes Iberia the ideal testing ground for the Substrate Isolation Criterion introduced in Chapter 2. If the criterion fails for Basque, it would be questionable for any less clear-cut case.

1.4.2 Central Europe: Extinct and Partially Attested Substrates

Central Europe presents a more challenging configuration. The pre-Roman substrates of the Alpine arc and the Po plain — Etruscan, Raetic, Venetic, and Illyrian — are extinct and, in the case of Raetic, poorly attested. The Romanisation of the region was thorough and early: most Central European toponymy has been filtered through Latin and subsequently through medieval vernacular languages. The phonotactic signal of the pre-Roman substrates must therefore be sought in the residual distributional properties of names that have survived this double filtering. The strong phonotactic distinctiveness of the Etruscan zone (the lowest bigram entropy in the Central European corpus, at 7.308 bits) confirms that substrate signal persists even under heavy Latinisation.

1.4.3 British Isles: Overlapping Celtic Varieties

The British Isles present the methodological challenge of phonotactically *similar* substrates: Brittonic and Goidelic Celtic are genetically related, sharing a common Proto-Celtic ancestor, and their phonotactic profiles are consequently much less divergent from one another than, say, Basque is from Romance. The framework's ability to discriminate between related substrate varieties is tested here. The Norse overlay in Northern Scotland and the Northern Isles, combined with the partial-attestation situation of Pictish, introduces additional complexity. Volume IV demonstrates that

the framework recovers the Celtic–Norse boundary with reasonable precision, and that Wales (Brittonic) is the most distinctive zone in the British corpus by the Substrate Isolation Criterion — a result that aligns with the well-known phonological conservatism of Welsh.

1.4.4 Indian Subcontinent: The Largest Corpus

The Indian Subcontinent presents the largest and most linguistically diverse corpus in the series ($N_{\text{total}} = 18{,}000$, drawn from over 700,000 available GeoNames entries for India and Pakistan). The primary substrate contrast is between Dravidian languages (concentrated in southern India) and Indo-Aryan languages (dominant in the north), with the Munda (Austro-Asiatic) family adding a third distinct phonotactic signature in central and eastern India. The Pakistani corpus introduces an additional contrast: the Northwest Frontier and Punjab zones reflect a complex overlay of Dardic, Iranian, and Turkic phonotactic patterns on a deeper Indo-Aryan base. The extraordinary phonotactic diversity of this region — the between-zone KL divergences are among the largest in the series — makes it both the most informative and the most complex volume.

1.4.5 Anatolia and the Caucasus: Maximum Stratification

Volume VI examines the region of highest substrate density in the series: Anatolia and the Caucasus, where documented linguistic stratification spans more than four millennia, from pre-Indo-European Hattic and Hurrian (second millennium BCE) through Greek, Armenian, and medieval Turkic layers to the modern configuration. The Kartvelian (Georgian) family, the Nakh-Daghestanian family (Chechen, Avar, Lak, and dozens of others), and the Northwest Caucasian family (Abkhaz, Circassian) are indigenous to the Caucasus with no demonstrated external genetic affiliations, making them, in principle, as informationally isolating as Basque in Iberia. Volume VI shows that the Armenian zone achieves the highest Bayesian Isolation Score in the entire series, confirming that the dense consonantism of Armenian phonology leaves a robust phonotactic footprint in toponymy.

1.5 Overview of the Series

Table 1.1 summarises the volumes in this series, together with the primary substrate contrasts analysed in each.

Table 1.1: Structure of the Toponomastic Entropy series.

Vol.	**Region**	**Primary substrates**	**Corpus size**
I	Methodological foundations	Iberian demonstration	13,896
II	Iberian Peninsula	Basque, Arabic, Romance, Arabic	44,390
III	Central Europe	Etruscan, Raetic, Venetic, Celtic	18,000
IV	British Isles	Pictish, Brythonic, Goidelic, Norse	13,005
V	Indian Subcontinent	Dravidian, Munda, Indo-Aryan	18,000
VI	Anatolia & Caucasus	Hattic, Hurrian, Urartian, Greek	18,000

Chapter 2

Mathematical Foundations

2.1 Probability Distributions over n-Gram Space

Definition: Character Alphabet and n-gram Space

A *character alphabet* Σ is a finite set of symbols. For the present series, Σ consists of the 26 lower-case Latin letters $\{a, b, \ldots, z\}$ plus a word-boundary marker #, giving $|\Sigma| = 27$. The *n-gram space* Σ^n is the set of all strings of length exactly n over Σ, with $|\Sigma^n| = 27^n$.

Definition: Toponymic Corpus

A *toponymic corpus* $\mathcal{T}$ for a geographic region $\mathcal{R}$ is a finite multiset of normalised toponym strings $\mathcal{T} = \{w_1, w_2, \ldots, w_N\} \subset \Sigma^*$, where each w_i is a normalised (lower-cased, diacritic-stripped) place name attested within $\mathcal{R}$, and $N = |\mathcal{T}|$ is the corpus size.

Definition: Character n-gram

Given $w \in \Sigma^*$ of length $|w| \geq n$, the multiset of *character n-grams* of w is

$$\mathcal{G}_n(w) \;=\; \big\{ w[i{:}i{+}n] \;\big|\; 0 \leq i \leq |w| - n \big\},$$

where $w[i{:}i{+}n]$ is the substring from position i to $i + n - 1$. Word-boundary markers are inserted at the start and end of w before extraction, so that `#w` and `w#` are valid bigrams for any toponym w.

For a corpus $\mathcal{T}$, the aggregate n-gram count function is

$$c_n(g; \mathcal{T}) \;=\; \sum_{w \in \mathcal{T}} \#\{g \in \mathcal{G}_n(w)\}, \quad g \in \Sigma^n.$$

The *empirical n-gram distribution* with Laplace smoothing parameter $\varepsilon > 0$ is

$$\hat{p}_n(g; \mathcal{T}) \;=\; \frac{c_n(g; \mathcal{T}) + \varepsilon}{\sum_{g' \in \Sigma^n} \left[c_n(g'; \mathcal{T}) + \varepsilon \right]}.$$

For comparative analyses across corpora, $\varepsilon = 10^{-9}$ (essentially zero smoothing) is used; for corpus-level entropy estimation, $\varepsilon = 1$ (Laplace smoothing) is preferred to avoid $\log(0)$ singularities.

2.2 Shannon Entropy of a Toponymic Corpus

Definition: Phonotactic Entropy

The *phonotactic entropy of order* n for a corpus $\mathcal{T}$ is

$$H_n(\mathcal{T}) \;=\; -\sum_{g \in \Sigma^n} \hat{p}_n(g;\mathcal{T}) \, \log_2 \hat{p}_n(g;\mathcal{T}).$$

It is expressed in bits and attains its maximum $n \log_2 |\Sigma|$ when all n-grams are equally frequent (maximum disorder) and its minimum 0 when all probability mass is concentrated on a single n-gram (complete order).

Proposition: Monotonicity and Corpus Size

For fixed alphabet Σ, n-gram order n, and increasing corpus size $N = |\mathcal{T}|$, the estimator $H_n(\mathcal{T})$ converges in probability from below to the true distribution entropy $H_n(p)$ as rare n-grams become observable. Specifically,

$$\mathbb{E}[H_n(\mathcal{T})] \;\leq\; H_n(p),$$

with the gap closing as $N \to \infty$. Subsampling to equal corpus size before comparing H_n values across corpora is therefore essential for fair comparison.

Proof. The bias of the plug-in entropy estimator is a classical result (Miller, 1955; Paninski, 2003). Let $p_g = p(g)$ be the true probability of n-gram g, $\hat{p}_g = c_n(g)/(\sum c_n)$ the empirical estimate. By the concavity of $-x \log x$ and Jensen's inequality, $\mathbb{E}[-\hat{p}_g \log \hat{p}_g] \leq -p_g \log p_g$, with equality only when $\hat{p}_g = p_g$ almost surely. Summing over g gives $\mathbb{E}[H_n(\mathcal{T})] \leq H_n(p)$. □

Remark: Entropy and Substrate Diversity

High phonotactic entropy indicates a corpus with diverse, nearly uniform n-gram usage: typical of a language with rich phonological combinations and few strong sequential constraints, or of a corpus containing a mixture of several distinct substrate layers (mixing increases entropy). Low entropy indicates concentration on a small set of highly recurrent patterns: typical of a recently standardised or phonologically simplified system. The contrast is directly observable in the Iberian data, where the Basque corpus attains $H_2 = 7.652$ bits against Castilian's 7.318 bits.

2.3 Conditional Entropy and the Chain Rule

The order-n phonotactic entropy H_n quantifies the overall disorder of the n-gram distribution, but it does not by itself separate two distinct sources of that disorder: the diversity of the underlying character inventory (captured by the unigram entropy H_1) and the strength of sequential dependencies between characters. *Conditional entropy* isolates the second component and is the natural information-theoretic counterpart to the autocorrelation structure of a symbol sequence.

Definition: Conditional Entropy of a Pair of Random Variables

Let A and B be discrete random variables with joint distribution $p(a, b)$ and marginals $p(a) = \sum_b p(a, b)$, $p(b) = \sum_a p(a, b)$. The *conditional entropy* of B given A is

$$H(B \mid A) \;=\; -\sum_{a} p(a) \sum_{b} p(b \mid a) \log_2 p(b \mid a) \;=\; \sum_{a} p(a)\, H(B \mid A = a),$$

where $p(b \mid a) = p(a, b)/p(a)$ and $H(B \mid A = a)$ is the Shannon entropy of the conditional distribution $p(\cdot \mid a)$. Equivalently, $H(B \mid A)$ is the expectation, over the marginal distribution of A, of the entropy remaining in B once the value of A is known.

Proposition: Chain Rule for Entropy

For any joint distribution $p(a, b)$,

$$H(A, B) \;=\; H(A) + H(B \mid A) \;=\; H(B) + H(A \mid B),$$

where $H(A, B) = -\sum_{a,b} p(a, b) \log_2 p(a, b)$ is the joint entropy of (A, B).

Proof. Write $p(a, b) = p(a)\, p(b \mid a)$ and expand the joint entropy:

$$\begin{aligned} H(A, B) &= -\sum_{a,b} p(a, b) \log_2 p(a, b) \\ &= -\sum_{a,b} p(a, b) \big[\log_2 p(a) + \log_2 p(b \mid a)\big] \\ &= -\sum_{a} p(a) \log_2 p(a) \sum_{b} p(b \mid a) - \sum_{a,b} p(a, b) \log_2 p(b \mid a) \\ &= H(A) + H(B \mid A), \end{aligned}$$

using $\sum_b p(b \mid a) = 1$ in the third line. The second equality follows by the symmetric argument with the roles of A and B exchanged. □

Applying the chain rule to consecutive characters of a toponymic corpus, where A ranges over Σ (the character preceding a boundary) and B over Σ (the character following it), and identifying the bigram distribution $\hat{p}_2$ with the joint distribution $p(a, b)$ and the unigram distribution $\hat{p}_1$ with the marginal $p(a)$, the chain rule specialises to the identity used throughout this series:

$$H_2(\mathcal{T}) \;=\; H_1(\mathcal{T}) + H(B \mid A), \qquad \text{i.e.} \qquad H(B \mid A) \;=\; H_2(\mathcal{T}) - H_1(\mathcal{T}).$$

This is the defining relation used for the conditional bigram entropy column reported in Chapter 11 (Table 11.4). It follows directly from Proposition 2.3 applied to the empirical bigram and unigram distributions, with no additional assumptions required.

Proposition: Bounds on Conditional Entropy

For any joint distribution $p(a, b)$ on a finite alphabet,

$$0 \;\leq\; H(B \mid A) \;\leq\; H(B),$$

with $H(B \mid A) = H(B)$ iff A and B are independent, and $H(B \mid A) = 0$ iff B is a deterministic function of A.

Proof. Non-negativity follows because $H(B \mid A) = \sum_a p(a)\, H(B \mid A = a)$ is a convex combination of entropies of the conditional distributions $p(\cdot \mid a)$, each of which is non-negative. The upper bound $H(B \mid A) \leq H(B)$ is equivalent, by the chain rule, to $H(A, B) \leq H(A) + H(B)$, which is itself equivalent to non-negativity of the mutual information $I(A; B) = H(B) - H(B \mid A) \geq 0$ (Gibbs' inequality applied to the joint versus the product-of-marginals distribution). Equality $H(B \mid A) = H(B)$ holds iff $I(A; B) = 0$ iff $p(a, b) = p(a)p(b)$ for all a, b, i.e. independence. The deterministic case follows because $H(B \mid A = a) = 0$ for every a with $p(a) > 0$ iff $p(\cdot \mid a)$ is degenerate for every such a, i.e. B is a function of A. □

Corollary: Conditional Entropy and Mutual Information

Combining the chain rule with the definition of mutual information yields the identity

$$I(A; B) \;=\; H(B) - H(B \mid A) \;=\; H(A) - H(A \mid B) \;=\; H(A) + H(B) - H(A, B).$$

In the toponymic setting, $I(A; B) = H_1(\mathcal{T}) - H(B \mid A) = 2H_1(\mathcal{T}) - H_2(\mathcal{T})$ measures the information that the identity of one character carries about the identity of its successor: a direct, corpus-level measure of positional phonotactic constraint, complementary to the conditional entropy itself.

Remark: Interpreting Conditional Entropy in the Toponymic Setting

Proposition 2.3 gives conditional bigram entropy a clean interpretive range for substrate analysis: at one extreme, $H(B \mid A) = H_1(\mathcal{T})$ would indicate that successive characters are statistically independent, so that the bigram distribution carries no information beyond what the unigram distribution already supplies; at the other extreme, $H(B \mid A) = 0$ would indicate a corpus in which every character is fully determined by its predecessor (an unrealistic limiting case for any natural-language toponym corpus). Empirically, every zone in the series falls strictly between these extremes, with $H(B \mid A)$ substantially below $H_1(\mathcal{T}) + \log_2 |\Sigma|$ but nowhere near zero, confirming the presence of genuine but partial phonotactic constraint rather than either pure randomness or rigid determinism at the character level.

The chain rule generalises beyond a single conditioning step. For an order-n Markov approximation of the toponym-generating process, the *entropy rate* is defined

as the limiting conditional entropy of a character given an unboundedly long preceding context:

$$H_\infty(\mathcal{T}) \;=\; \lim_{n\to\infty} H\big(X_n \mid X_1, \ldots, X_{n-1}\big),$$

a quantity that decreases monotonically as n increases (conditioning can only reduce or preserve entropy, by Proposition 2.3 applied iteratively) and that is bounded below by zero for any finite alphabet. The present series approximates $H_\infty(\mathcal{T})$ from below using $n \le 3$; the bigram conditional entropy $H(B \mid A) = H_2 - H_1$ is the first-order approximation in this hierarchy, and the trigram conditional entropy $H_3 - H_2$ (used implicitly in the trigram-versus-bigram comparison of Chapter 11) is the second.

2.4 Perplexity and Its Relation to Cross-Entropy

Held-out perplexity, used in Chapter 5 and Chapter 11 as a model-validation device, is a monotonic transform of cross-entropy and is recorded here formally for completeness.

Definition: Perplexity

Let $\hat{q}$ be an n-gram model (a probability distribution over Σ^n) fitted to a training corpus, and let w be a held-out toponym with n-gram multiset $\mathcal{G}_n(w)$ of size $|w|+1$ (including boundary markers). The *perplexity* of w under $\hat{q}$ is

$$\mathrm{PP}_{\hat{q}}(w) \;=\; 2^{\,H_{\times w}(\hat{q})}, \qquad H_{\times w}(\hat{q}) \;=\; -\frac{1}{|\mathcal{G}_n(w)|} \sum_{g \in \mathcal{G}_n(w)} \log_2 \hat{q}(g),$$

i.e. perplexity is two raised to the average per-n-gram cross-entropy of w under the model $\hat{q}$.

Proposition: Perplexity as Effective Vocabulary Size

If $\hat{q}$ is the uniform distribution over a support of size V, then $\mathrm{PP}_{\hat{q}}(w) = V$ for every w whose n-grams lie within that support. More generally, for any $\hat{q}$, $\mathrm{PP}_{\hat{q}}(w) \le |\Sigma|^n$, with equality only when $\hat{q}$ is uniform over Σ^n and all n-grams of w are distinct.

Proof. For uniform $\hat{q}$ with support size V, $\hat{q}(g) = 1/V$ for every g in the support, so $H_{\times w}(\hat{q}) = \log_2 V$ and $\mathrm{PP}_{\hat{q}}(w) = 2^{\log_2 V} = V$ directly from Definition 2.4. For the general bound, note that $\hat{q}(g) \le 1$ for every g, so $-\log_2 \hat{q}(g) \ge 0$; combined with the fact that the uniform distribution over Σ^n maximises the per-symbol cross-entropy averaged over any fixed empirical distribution of n-grams (a consequence of the non-negativity of D_{KL} between the empirical n-gram frequencies of w and $\hat{q}$), the cross-entropy $H_{\times w}(\hat{q})$ is maximised, for fixed w, at $\log_2 |\Sigma|^n = n \log_2 |\Sigma|$, giving the stated bound. □

This proposition motivates the informal reading of perplexity as the size of an equivalent uniform vocabulary: a model that assigns perplexity 11 to a held-out toponym is, on average, as uncertain about each n-gram as a model that chooses

uniformly among 11 equally likely alternatives. This is the basis for the "self/cross perplexity ratio" diagnostic of Chapter 11: a ratio near 1 indicates that a foreign-zone model is, in this effective-vocabulary sense, almost as informative as the zone's own model, whereas a ratio well above 1 indicates a substantial loss of effective predictive power when the wrong model is applied.

Corollary: Perplexity and KL Divergence

For a held-out corpus $\mathcal{T}'$ drawn from the same distribution $\hat{p}$ used to define the empirical n-gram frequencies, the expected log-perplexity under a model $\hat{q}$ decomposes as

$$\mathbb{E}_{\hat{p}}\big[\log_2 \mathrm{PP}_{\hat{q}}(W)\big] \;=\; H_{\times}(\hat{p},\hat{q}) \;=\; H_n(\hat{p}) + D_{\mathrm{KL}}(\hat{p}\,\|\,\hat{q}),$$

by the cross-entropy identity established below (Section 2.5). Consequently, perplexity is minimised, over all candidate models $\hat{q}$, by $\hat{q} = \hat{p}$, and the excess log-perplexity of any other model is exactly the KL divergence from the true distribution to the candidate model.

Corollary 2.4 clarifies why self-perplexity (model and held-out data drawn from the same zone) is, in expectation, the lowest perplexity achievable for that zone, and why the gap to any cross-volume model's perplexity is a monotonic function of the KL divergence between the two zones' phonotactic distributions. This is the theoretical basis for using the self/cross perplexity ratio as a proxy for phonotactic distinctiveness, parallel to but methodologically independent of the KL-divergence-based Bayesian Isolation Score of Theorem 2.8: where the BIS is computed directly from the fitted training distributions, the perplexity ratio is computed from out-of-sample predictive performance, and the two measures agreeing in their ranking of zones (Chapter 11) constitutes a form of external validation of the BIS via a distinct, held-out-data-based criterion.

2.5 Kullback–Leibler Divergence and Related Measures

Definition: Kullback–Leibler Divergence

Let $\hat{p}$ and $\hat{q}$ be the n-gram distributions of corpora $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively. The *Kullback–Leibler divergence* (or *relative entropy*) of $\hat{p}$ from $\hat{q}$ is

$$D_{\mathrm{KL}}(\hat{p}\|\hat{q}) \;=\; \sum_{g\in\Sigma^n} \hat{p}(g)\,\log_2 \frac{\hat{p}(g)}{\hat{q}(g)}.$$

Proposition: Fundamental Properties of KL Divergence

The KL divergence satisfies the following:

(i) $D_{\mathrm{KL}}(\hat{p}\|\hat{q}) \geq 0$ for all distributions, with equality iff $\hat{p} = \hat{q}$ almost everywhere (*Gibbs' inequality*).

(ii) $D_{\mathrm{KL}}(\hat{p}\|\hat{q}) \neq D_{\mathrm{KL}}(\hat{q}\|\hat{p})$ in general (asymmetry). The two directions have distinct interpretations: $D_{\mathrm{KL}}(\hat{p}\|\hat{q})$ is the expected excess code length when $\hat{q}$-optimal codes are used to encode $\hat{p}$-distributed symbols; $D_{\mathrm{KL}}(\hat{q}\|\hat{p})$ is the converse.

(iii) $D_{\mathrm{KL}}(\hat{p}\|\hat{q}) \leq \log_2 \frac{1}{\hat{q}(g^*)}$ for the g^* minimising $\hat{q}(g^*)$, providing an upper bound in terms of the rarest n-gram under the reference distribution.

(iv) The symmetrised Jensen–Shannon divergence $\mathrm{JS}(\hat{p},\hat{q}) = \frac{1}{2}D_{\mathrm{KL}}(\hat{p}\|\hat{m}) + \frac{1}{2}D_{\mathrm{KL}}(\hat{q}\|\hat{m})$, where $\hat{m} = \frac{\hat{p}+\hat{q}}{2}$, is a proper metric on probability simplices (Endres and Schindelin, 2003).

Definition: Cross-Entropy

The *cross-entropy* of $\hat{p}$ with respect to $\hat{q}$ is

$$H_{\times}(\hat{p},\hat{q}) = -\sum_{g\in\Sigma^n} \hat{p}(g)\,\log_2 \hat{q}(g) = H_n(\mathcal{T}_1) + D_{\mathrm{KL}}(\hat{p}\|\hat{q}).$$

The cross-entropy matrix $[H_{\times}(\hat{p}_i,\hat{p}_j)]_{i,j}$ provides a comprehensive picture of pairwise phonotactic distances between substrate zones. Its diagonal entries equal the self-entropies $H_n(\mathcal{T}_i)$; off-diagonal asymmetries reveal directional phonotactic relationships.

Definition: Symmetric Phonotactic Distance

The *symmetric phonotactic distance* between corpora $\mathcal{T}_i$ and $\mathcal{T}_j$ is

$$\delta(\mathcal{T}_i,\mathcal{T}_j) = \tfrac{1}{2}\big[D_{\mathrm{KL}}(\hat{p}_i\|\hat{p}_j) + D_{\mathrm{KL}}(\hat{p}_j\|\hat{p}_i)\big].$$

This is not a metric (it fails the triangle inequality in general) but satisfies non-negativity and symmetry, making it suitable for use as a dissimilarity in clustering algorithms.

2.6 Jensen–Shannon Divergence as a Metric

The Jensen–Shannon divergence (JSD) is of special importance in the present series because, unlike KL divergence, it is bounded, symmetric, and — when defined with the square root — a proper metric on the probability simplex. This section provides a complete treatment of its properties and the proof that $\sqrt{\mathrm{JS}}$ is a metric.

Definition: Jensen–Shannon Divergence

Let $\hat{p}$ and $\hat{q}$ be n-gram distributions for corpora $\mathcal{T}_1$ and $\mathcal{T}_2$. The *Jensen–Shannon divergence* is

$$\mathrm{JS}(\hat{p},\hat{q}) = \frac{1}{2}D_{\mathrm{KL}}(\hat{p}\|\hat{m}) + \frac{1}{2}D_{\mathrm{KL}}(\hat{q}\|\hat{m}), \quad \hat{m} = \frac{\hat{p}+\hat{q}}{2}.$$

$\mathrm{JS}(\hat{p}, \hat{q}) \in [0, 1]$ when computed in bits ($\log_2$), attaining 0 iff $\hat{p} = \hat{q}$ and 1 iff $\hat{p}$ and $\hat{q}$ have disjoint support.

Proposition: Boundedness and Symmetry of JSD

For any distributions $\hat{p}$, $\hat{q}$:

(i) $0 \leq \mathrm{JS}(\hat{p}, \hat{q}) \leq 1$ (in bits);

(ii) $\mathrm{JS}(\hat{p}, \hat{q}) = \mathrm{JS}(\hat{q}, \hat{p})$;

(iii) $\mathrm{JS}(\hat{p}, \hat{q}) = H(\hat{m}) - \frac{H(\hat{p})+H(\hat{q})}{2}$.

Proof. (i) The lower bound JS ≥ 0 follows from the non-negativity of KL divergence (Gibbs' inequality). For the upper bound, note that $D_{\mathrm{KL}}(\hat{p}\|\hat{m}) \leq \log_2(1/\min_g \hat{m}(g)) \leq \log_2 2 = 1$ bit when $\hat{p}$ and $\hat{q}$ have disjoint support and $\hat{m}(g) = \hat{p}(g)/2$ or $\hat{q}(g)/2$ on each g; more precisely, $D_{\mathrm{KL}}(\hat{p}\|\hat{m}) = \sum_g \hat{p}(g) \log_2(2\hat{p}(g)/(\hat{p}(g) + \hat{q}(g))) \leq \sum_g \hat{p}(g) \log_2 2 = 1$ since $\hat{p}(g) + \hat{q}(g) \geq \hat{p}(g)$. By symmetry, $D_{\mathrm{KL}}(\hat{q}\|\hat{m}) \leq 1$, so JS ≤ 1. (ii) The mixture $\hat{m}$ is symmetric in $\hat{p}$ and $\hat{q}$. (iii) Expand: $H(\hat{m}) = -\sum_g \hat{m}(g) \log_2 \hat{m}(g)$. Then $D_{\mathrm{KL}}(\hat{p}\|\hat{m}) = -H(\hat{p}) - \sum_g \hat{p}(g) \log_2 \hat{m}(g)$, and similarly for $\hat{q}$. Adding the two and dividing by 2 gives $\mathrm{JS}(\hat{p}, \hat{q}) = H(\hat{m}) - (H(\hat{p}) + H(\hat{q}))/2$. □

Theorem: $\sqrt{\mathrm{JS}}$ is a Metric (,)

The function $d_{\mathrm{JS}}(\hat{p}, \hat{q}) = \sqrt{\mathrm{JS}(\hat{p}, \hat{q})}$ satisfies the four axioms of a metric on the probability simplex: non-negativity, identity of indiscernibles, symmetry, and the triangle inequality.

Proof. Non-negativity and symmetry follow from the corresponding properties of JSD. Identity: $d_{\mathrm{JS}} = 0$ iff JS $= 0$ iff $\hat{p} = \hat{q}$. For the triangle inequality, consider distributions $\hat{p}, \hat{q}, \hat{r}$. Define $\hat{m}_{pr} = (\hat{p} + \hat{r})/2$, $\hat{m}_{pq} = (\hat{p} + \hat{q})/2$, $\hat{m}_{qr} = (\hat{q} + \hat{r})/2$. The key lemma is that H is concave, so $H(\hat{m}_{pr}) \leq (H(\hat{m}_{pq}) + H(\hat{m}_{qr}))/2 + H((\hat{m}_{pq} + \hat{m}_{qr})/2)/2$. Since $(\hat{m}_{pq} + \hat{m}_{qr})/2 = (\hat{p} + 2\hat{q} + \hat{r})/4 \neq \hat{m}_{pr}$ in general, the proof proceeds by the embedding of the probability simplex into a Hilbert space (the Hellinger space) and the Cauchy–Schwarz inequality; see Endres and Schindelin (2003) for the complete argument. The result is $d_{\mathrm{JS}}(\hat{p}, \hat{r}) \leq d_{\mathrm{JS}}(\hat{p}, \hat{q}) + d_{\mathrm{JS}}(\hat{q}, \hat{r})$. □

The metric property of d_{JS} is practically significant: it means that the pairwise JSD matrix $[d_{\mathrm{JS}}(\hat{p}_i, \hat{p}_j)]$ can be used directly as a distance matrix in metric multidimensional scaling (MDS) and in agglomerative hierarchical clustering with Ward linkage, both of which are formally defined only for proper metrics.

Remark: JSD vs. Symmetric KL in Practice

In all volumes of this series, the symmetric phonotactic distance $\delta(\mathcal{T}_i, \mathcal{T}_j) = [D_{\mathrm{KL}}(\hat{p}_i\|\hat{p}_j) + D_{\mathrm{KL}}(\hat{p}_j\|\hat{p}_i)]/2$ is used for numerical reporting, as it provides larger absolute values (easier to compare across regions) and is scale-free with respect to corpus size at leading order. The JSD matrix is used for hierarchical cluster-

ing and for the dendrogram visualisations in Chapters 6 and 11. Both measures identify the same pairs of zones as most and least similar in every volume.

2.7 Rényi Entropy and Its Role as a Generalisation

Shannon entropy occupies a privileged position among information measures because of its unique axiomatic characterisation (non-negativity, continuity, maximality at the uniform distribution, and the chain rule). However, a one-parameter family of generalisations, due to Rényi (1961), is of direct relevance to the present series: Rényi entropy of order α provides a continuous deformation of Shannon entropy that emphasises either rare n-grams (for $\alpha < 1$) or frequent ones (for $\alpha > 1$), and thus allows a more nuanced characterisation of the tails of the phonotactic distribution.

Definition: Rényi Entropy of Order α

Let $\hat{p}$ be an n-gram distribution over Σ^n. The *Rényi entropy of order* $\alpha \geq 0$, $\alpha \neq 1$ is

$$H_\alpha(\hat{p}) \;=\; \frac{1}{1-\alpha}\,\log_2\!\left(\sum_{g\in\Sigma^n}\hat{p}(g)^\alpha\right).$$

The limiting cases are:

$$\lim_{\alpha\to 0} H_\alpha(\hat{p}) = \log_2 |\mathrm{supp}(\hat{p})|, \quad \text{(Hartley entropy: counts support size)}$$

$$\lim_{\alpha\to 1} H_\alpha(\hat{p}) = -\sum_g \hat{p}(g)\log_2\hat{p}(g) = H_1(\hat{p}), \quad \text{(Shannon entropy)}$$

$$\lim_{\alpha\to\infty} H_\alpha(\hat{p}) = -\log_2 \max_g \hat{p}(g), \quad \text{(min-entropy: focuses on the dominant } n\text{-gram)}.$$

Proposition: Monotonicity of Rényi Entropy in α

For any distribution $\hat{p}$ on a finite alphabet, $H_\alpha(\hat{p})$ is a non-increasing function of α. That is,

$$\alpha_1 < \alpha_2 \implies H_{\alpha_1}(\hat{p}) \;\geq\; H_{\alpha_2}(\hat{p}),$$

with equality iff $\hat{p}$ is the uniform distribution.

Proof. Write $f(\alpha) = (1-\alpha)^{-1}\log_2\sum_g \hat{p}(g)^\alpha$. Taking $\frac{d}{d\alpha}f(\alpha)$ and examining the sign requires verifying that $\frac{d}{d\alpha}\log\sum_g \hat{p}(g)^\alpha = \frac{\sum_g \hat{p}(g)^\alpha \log\hat{p}(g)}{\sum_g \hat{p}(g)^\alpha}$, i.e. the $\hat{p}^\alpha$-weighted average of $\log\hat{p}(g)$, is a decreasing function of $(1-\alpha)^{-1}$. By the log-sum inequality and the convexity of $x \mapsto x^\alpha$ for $\alpha > 1$ (resp. concavity for $\alpha \in (0,1)$), the weighted average $\sum_g \frac{\hat{p}(g)^\alpha}{\sum_{g'}\hat{p}(g')^\alpha}\log\hat{p}(g)$ decreases as α increases (it shifts weight toward the highest-probability n-grams, which have the least negative $\log\hat{p}$, but the denominator grows at the same rate in the uniform case). The complete formal proof is by the Hölder inequality; see, e.g., van Erven and Harremos (2014) for a detailed treatment. □

Corollary: Rényi Entropy as an Inequality Detector

The gap $H_0(\hat{p}) - H_\infty(\hat{p})$ quantifies the *unevenness* of $\hat{p}$: it is zero iff $\hat{p}$ is uniform over its support, and positive otherwise, with larger values indicating greater concentration of the distribution on a small subset of frequent n-grams. For the toponymic setting, the Rényi gap

$$\Delta_{\mathrm{R}}(\mathcal{T}) = H_0(\hat{p}) - H_2(\hat{p})$$

(taking $\alpha = 2$ for the collision entropy) measures how far the corpus is from the combinatorially available maximum: large Δ_{R} indicates a corpus dominated by a small set of high-frequency n-grams atop a long tail of rare ones.

The order-$\alpha = 2$ Rényi entropy, $H_2(\hat{p}) = -\log_2 \sum_g \hat{p}(g)^2$, is also known as the *collision entropy*: it equals $-\log_2 P_c$, where $P_c = \sum_g \hat{p}(g)^2$ is the probability that two independently and identically distributed samples from $\hat{p}$ produce the same n-gram. This quantity is directly estimable from corpus data without knowledge of the full distribution, requiring only a count of coincident bigrams in a resampled pair of subcorpora.

Proposition: Relation Between Shannon and Rényi Entropies in the Toponymic Setting

For the bigram distributions $\hat{p}_2(\cdot;\mathcal{T})$ over Σ^2 (alphabet size $K = |\Sigma|^2 = 729$ in the 27-character system), the following inequalities hold:

(i) $H_\alpha(\hat{p}) \le \log_2 K$ for all $\alpha \ge 0$, with equality iff $\hat{p}$ is the uniform distribution.

(ii) $H_2(\hat{p}) \le H_1(\hat{p})$ (collision entropy does not exceed Shannon entropy for any distribution).

(iii) The gap $H_1(\hat{p}) - H_2(\hat{p})$ is a measure of the *dispersion* of $\hat{p}$ beyond what a uniform distribution over the same support would exhibit; it vanishes at uniformity and increases with the concentration of probability mass on high-frequency bigrams.

Proof. Part (i) follows from the monotonicity of Rényi entropy and the fact that $H_0(\hat{p}) = \log_2 |\mathrm{supp}(\hat{p})| \le \log_2 K$. Part (ii) is an immediate consequence of Proposition 2.7 with $\alpha_1 = 1$, $\alpha_2 = 2$. Part (iii): the gap $H_1 - H_2 = -\sum_g \hat{p}(g) \log_2 \hat{p}(g) + \log_2 \sum_g \hat{p}(g)^2$; Jensen's inequality applied to $-\log_2$ (a convex function) with the weight vector $(\hat{p}(g))_g$ gives $-\log_2 \sum_g \hat{p}(g)^2 \le -\sum_g \hat{p}(g) \log_2 \hat{p}(g)$, confirming $H_1 \ge H_2$; equality holds iff $\hat{p}(g) = \hat{p}(g)^2$ for all g, i.e. iff $\hat{p}$ is supported on a single n-gram. □

Remark: Rényi Entropy in the Iberian Data: Illustrative Comparison

For the Iberian corpus (real GeoNames data), the Basque zone has Shannon bigram entropy $H_1(\hat{p}_2) = 7.652$ bits and collision entropy $H_2(\hat{p}_2) \approx 7.33$ bits (computed from the collision probability $P_c = \sum_g \hat{p}_2(g)^2$), giving a Rényi gap of approximately 0.32 bits. The Castilian zone has $H_1 = 7.318$ bits and $H_2 \approx 6.94$ bits,

giving a Rényi gap of ≈ 0.38 bits. The larger Castilian Rényi gap reflects the higher concentration of the Castilian bigram distribution on a small set of dominant transitions (`ar`, `al`, `os`), consistent with the stronger consonant-vowel alternation constraint already noted in the spectral analysis of the transition matrix (Section 6.2). No Rényi entropy values are used as primary evidence in any volume of the series; they are reported here as a supplementary characterisation of distributional shape.

Definition: Rényi Divergence

The *Rényi divergence* of order $\alpha > 0$, $\alpha \neq 1$, from $\hat{q}$ to $\hat{p}$ is

$$D_\alpha(\hat{p}\|\hat{q}) \;=\; \frac{1}{\alpha - 1}\,\log_2\!\left(\sum_g \hat{p}(g)^\alpha\,\hat{q}(g)^{1-\alpha}\right).$$

The limiting case $\alpha \to 1$ recovers the Kullback–Leibler divergence: $D_1(\hat{p}\|\hat{q}) = D_{\mathrm{KL}}(\hat{p}\|\hat{q})$. For $\alpha = 1/2$, the Rényi divergence is the *Bhattacharyya divergence* $D_{1/2} = -2\log_2 \sum_g \sqrt{\hat{p}(g)\hat{q}(g)}$, whose exponential $e^{-D_{1/2}}$ is the *Bhattacharyya coefficient*, measuring the overlap between the two distributions.

The Bhattacharyya coefficient provides a scale-bounded (in $[0, 1]$) measure of overlap between phonotactic distributions that complements the KL divergence. Because it is symmetric in $\hat{p}$ and $\hat{q}$ and bounded, it is well-suited for visualisation: a heatmap of pairwise Bhattacharyya coefficients provides an immediately interpretable picture of which zone pairs share the most phonotactic material, without the unbounded scale and asymmetry of the KL heatmaps of Chapter 11. For the Iberian corpus, the Basque–Castilian Bhattacharyya coefficient is approximately 0.61, the Basque–Galician coefficient is 0.58, and the Castilian–Galician coefficient is 0.87, consistent in direction (if not in precise magnitude) with the KL divergence ordering.

Proposition: Rényi Divergence and the Data-Processing Inequality

For any deterministic function $f : \Sigma^n \to \mathcal{Y}$ (e.g. a coarsening of the n-gram space to a smaller feature alphabet), the induced distributions $f_*\hat{p}$ and $f_*\hat{q}$ (push-forwards under f) satisfy

$$D_\alpha(f_*\hat{p}\|f_*\hat{q}) \;\leq\; D_\alpha(\hat{p}\|\hat{q}) \quad \text{for all } \alpha \geq 0.$$

In particular, $D_{\mathrm{KL}}(f_*\hat{p}\|f_*\hat{q}) \leq D_{\mathrm{KL}}(\hat{p}\|\hat{q})$: any coarsening of the n-gram representation can only reduce or preserve the information-theoretic distance between two corpora, never increase it. This is the fundamental justification for the claim in Section 7.3 that zone-level analysis (f = zone label) captures at most as much spatial structure as the underlying continuous geographic representation.

Proof. The data-processing inequality for Rényi divergence follows from the log-sum inequality (a consequence of Jensen's inequality applied to $t \mapsto t^\alpha$ for $\alpha > 1$, resp. concavity for $\alpha < 1$). For f deterministic, $f_*\hat{p}(y) = \sum_{g:f(g)=y} \hat{p}(g)$ and similarly for

$\hat{q}$. The key step is $\sum_y f_*\hat{p}(y)^\alpha f_*\hat{q}(y)^{1-\alpha} = \sum_y \big(\sum_{g:f(g)=y} \hat{p}(g)\big)^\alpha \big(\sum_{g:f(g)=y} \hat{q}(g)\big)^{1-\alpha} \leq \sum_g \hat{p}(g)^\alpha \hat{q}(g)^{1-\alpha}$, where the inequality is the generalised Hölder inequality applied within each f-fibre. Taking log and multiplying by $(1-\alpha)^{-1}$ gives the result. □

2.8 The Mixing Entropy Lemma

A fundamental result motivating the entropy-based approach is that mixing two distributions increases entropy. This lemma underlies the entire rationale for substrate detection: a corpus drawn from two distinct substrate layers will exhibit higher entropy than either layer alone.

Theorem: Mixing Entropy Lemma

Let $\hat{p}$ and $\hat{q}$ be n-gram distributions with entropies $H(\hat{p})$ and $H(\hat{q})$, and let $\hat{m}_\lambda = \lambda\hat{p} + (1-\lambda)\hat{q}$ for $\lambda \in (0,1)$. Then

$$H(\hat{m}_\lambda) \;\geq\; \lambda\, H(\hat{p}) + (1-\lambda)\, H(\hat{q}),$$

with equality iff $\hat{p} = \hat{q}$.

Proof. $-x\log x$ is a strictly concave function of $x \geq 0$. By Jensen's inequality,

$$\begin{aligned} H(\hat{m}_\lambda) &= \sum_g \big[-\hat{m}_\lambda(g)\log\hat{m}_\lambda(g)\big] \\ &= \sum_g \big[-(\lambda\hat{p}(g) + (1-\lambda)\hat{q}(g))\log(\lambda\hat{p}(g) + (1-\lambda)\hat{q}(g))\big] \\ &\geq \sum_g \big[\lambda(-\hat{p}(g)\log\hat{p}(g)) + (1-\lambda)(-\hat{q}(g)\log\hat{q}(g))\big] \\ &= \lambda\, H(\hat{p}) + (1-\lambda)\, H(\hat{q}). \end{aligned}$$

Strict concavity gives equality iff $\hat{p}(g) = \hat{q}(g)$ for all g. □

Corollary: Entropy Inflation as a Substrate Signal

Suppose a toponymic corpus $\mathcal{T}$ is a mixture of k substrate layers $\mathcal{T}_1, \ldots, \mathcal{T}_k$ with mixing proportions $\pi_1, \ldots, \pi_k$ ($\sum_z \pi_z = 1$). Then

$$H_n(\mathcal{T}) \;\geq\; \sum_{z=1}^{k} \pi_z\, H_n(\mathcal{T}_z),$$

with equality only if all substrate n-gram distributions are identical. The *entropy inflation* $\Delta H = H_n(\mathcal{T}) - \sum_z \pi_z H_n(\mathcal{T}_z)$ measures the phonotactic heterogeneity of the corpus beyond what each component contributes individually. It equals the average JSD between the mixture and its components, weighted by the mixing proportions:

$$\Delta H \;=\; \sum_z \pi_z \cdot D_{\mathrm{KL}}(\hat{p}_z \| \hat{p}),$$

where $\hat{p} = \sum_z \pi_z \hat{p}_z$ is the pooled distribution.

The entropy inflation corollary provides a direct formula for interpreting the excess entropy of a pooled corpus: it is a weighted average of the KL divergences of each component from the pooled distribution, i.e. the average Bayesian Isolation Score over the corpus. In the Iberian prototype, $\Delta H_2 \approx 0.15$ bits, reflecting the modest but detectable phonotactic heterogeneity of the five-zone corpus.

Theorem: Substrate Isolation Criterion

Let $\{\mathcal{T}_1, \ldots, \mathcal{T}_k\}$ be k toponymic corpora from geographically adjacent zones. Zone i satisfies the *Substrate Isolation Criterion* at threshold $\delta > 1$ if

$$\frac{\min_{j \neq i} D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j)}{\max_{\substack{j,\ell \neq i \\ j \neq \ell}} D_{\mathrm{KL}}(\hat{p}_j \| \hat{p}_\ell)} > \delta.$$

When satisfied, zone i is informationally isolated from the reference group $\{\mathcal{T}_j : j \neq i\}$ at level δ.

The Substrate Isolation Criterion formalises the intuition that a linguistic isolate (a language with no known genetic relatives within the study area) will produce a phonotactic distribution that is *simultaneously* distant from all other distributions in the study, not merely from the closest one.

Corollary: Sensitivity to Corpus Size

The Substrate Isolation Criterion is sensitive to corpus size through the bias of the entropy estimator. At small corpus sizes ($N < 500$ per zone), the estimated KL divergences may underestimate the true divergences due to sparse n-gram coverage. At the corpus sizes used in this series ($N \geq 3{,}000$ per zone for all zones except Basque, which uses its full GeoNames pool of $N = 1{,}896$), simulation studies (Section 3.4) indicate that the bias is negligible for the principal conclusions.

Example: Basque Isolation in the Iberian Data

In the Iberian corpus (real GeoNames data, balanced sample), the KL divergences from the Basque zone to all other zones range from $D_{\mathrm{KL}}(\text{Basque} \| \text{Catalan}) = 0.776$ bits to $D_{\mathrm{KL}}(\text{Basque} \| \text{Castilian}) = 0.934$ bits, while the maximum pairwise KL divergence among the four non-Basque zones is $D_{\mathrm{KL}}(\text{Castilian} \| \text{Galician}) = 0.607$ bits. The isolation ratio is $0.776/0.607 = 1.278 > 1$, satisfying the Criterion at threshold $\delta = 1.278$. This result is consistent with the well-established pre-Indo-European status of the Basque language and constitutes the primary empirical validation of the Substrate Isolation Criterion on the real GeoNames corpus.

2.9 Mutual Information and Spatial Gradients

Beyond pairwise comparison, one may ask whether the phonotactic distribution varies *continuously* across space, suggesting gradual substrate mixing rather than

sharp areal boundaries. This question is formalised through spatial mutual information.

Definition: Spatial Mutual Information

Let $G \in \Sigma^n$ be an n-gram random variable and $\mathbf{x} = (\lambda, \phi)$ a geographic coordinate. The *spatial mutual information* is

$$I(G; \mathbf{x}) \;=\; \sum_{g} \int p(g, \mathbf{x}) \, \log_2 \frac{p(g, \mathbf{x})}{p(g)\, p(\mathbf{x})} \, \mathrm{d}\mathbf{x},$$

where $p(g, \mathbf{x})$ is estimated by kernel density estimation over the corpus with geographic coordinates attached to each toponym.

n-grams with high spatial mutual information are those whose frequency is strongly geographically structured — prime candidates for substrate-diagnostic features. This measure allows the construction of *substrate signature maps*: rasters of the peninsula or region colour-coded by the expected value of a given substrate-indicator n-gram, revealing the spatial extent of a substrate layer without pre-defining its boundaries.

Proposition: Decomposition of Phonotactic Variance

The total phonotactic variance of a region can be decomposed as:

$$H(\text{pooled}) - \bar{H}(\text{zones}) \;=\; \frac{1}{k} \sum_{i=1}^{k} I(\text{Zone}_i; G),$$

where $\bar{H}(\text{zones})$ is the mean within-zone entropy and $H(\text{pooled})$ is the entropy of the pooled corpus. The left-hand side measures the entropy inflation due to zone mixing; the right-hand side decomposes this inflation into per-zone contributions.

Chapter 3

Statistical Estimation and Hypothesis Testing

3.1 The Multinomial Sampling Model

The fundamental probabilistic model underlying the present framework treats the generation of toponyms as a two-stage process. First, a substrate layer $Z \in \{1, \ldots, k\}$ is selected according to the prior $\pi_z = P(Z = z)$; then, a toponym is generated by drawing characters sequentially from the n-gram transition model $p(\cdot|z)$ of layer z.

For estimation purposes, we abstract from the generative process and work directly with the marginal n-gram counts. For a zone z with corpus $\mathcal{T}_z$, the count vector $\mathbf{c}_z = (c_n(g; \mathcal{T}_z))_{g \in \Sigma^n}$ is modelled as a draw from a Multinomial distribution:

$$\mathbf{c}_z \sim \text{Mult}(N_z, \boldsymbol{\theta}_z),$$

where $N_z = \sum_g c_n(g; \mathcal{T}_z)$ is the total n-gram count in zone z and $\boldsymbol{\theta}_z = (\theta_{z,g})_{g \in \Sigma^n}$ is the true n-gram probability vector.

Definition: Maximum Likelihood Estimator

The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}_z$ under the Multinomial model is the relative frequency vector:

$$\hat{\theta}^{\text{MLE}}_{z,g} = \frac{c_n(g; \mathcal{T}_z)}{N_z}.$$

The MLE is unbiased but has zero probability for unobserved n-grams, making it unsuitable for KL divergence computation (which would yield infinite divergence). Laplace-smoothed estimates with $\varepsilon = 1$ are used throughout.

3.2 Confidence Intervals for Entropy Estimates

The bias and variance of the plug-in entropy estimator are well-studied (Miller, 1955; Grassberger, 1988; Paninski, 2003). For practical purposes, we use the following first-order correction:

Proposition: Miller–Madow Bias Correction

The *Miller–Madow corrected* entropy estimator is

$$\hat{H}_n^{\mathrm{MM}}(\mathcal{T}) \;=\; \hat{H}_n^{\mathrm{MLE}}(\mathcal{T}) + \frac{S-1}{2N_z \ln 2},$$

where $S = |\{g : c_n(g;\mathcal{T}) > 0\}|$ is the number of observed n-gram types and N_z is the total count. The correction term converges to zero as $N_z \to \infty$.

Bootstrap confidence intervals for H_n are computed by resampling toponyms with replacement from $\mathcal{T}$ and recomputing the entropy estimator. The 95% bootstrap CI is $[\hat{H}^{(0.025)}, \hat{H}^{(0.975)}]$ over 1,000 bootstrap replicates. At the corpus sizes used in this series, the 95% CI width for bigram entropy is approximately ± 0.008 bits, confirming that all inter-zone entropy differences reported are statistically significant.

3.2.1 Derivation of the Miller–Madow Correction

The correction term is obtained from a second-order Taylor expansion of the plug-in entropy around the true distribution $\boldsymbol{\theta}$. Write $\hat{\theta}_g = \theta_g + \delta_g$, where δ_g is the sampling fluctuation of the relative frequency at cell g, with $\mathbb{E}[\delta_g] = 0$ and $\mathrm{Var}(\delta_g) = \theta_g(1-\theta_g)/N$ under the Multinomial model. The plug-in entropy (in nats, for convenience) is

$$\hat{H}^{\mathrm{MLE}} = -\sum_g \hat{\theta}_g \ln \hat{\theta}_g = -\sum_g (\theta_g + \delta_g)\ln(\theta_g + \delta_g).$$

Expanding $\ln(\theta_g + \delta_g) = \ln\theta_g + \delta_g/\theta_g - \delta_g^2/(2\theta_g^2) + O(\delta_g^3)$ and multiplying through by $(\theta_g + \delta_g)$, the terms linear in δ_g vanish in expectation (since $\mathbb{E}[\delta_g] = 0$), and collecting the quadratic terms gives

$$\begin{aligned}
\mathbb{E}\big[\hat{H}^{\mathrm{MLE}}\big] - H &= -\frac{1}{2}\sum_g \frac{\mathrm{Var}(\delta_g)}{\theta_g} + O(N^{-2}) \\
&= -\frac{1}{2N}\sum_g (1-\theta_g) + O(N^{-2}) \\
&= -\frac{S^* - 1}{2N} + O(N^{-2}),
\end{aligned}$$

where $S^* = |\{g : \theta_g > 0\}|$ is the true (population) support size, since $\sum_g (1-\theta_g) = S^* - 1$ over the support. The plug-in estimator is therefore negatively biased by approximately $(S^*-1)/(2N)$ nats. Substituting the empirical support size S for the (unobservable) true support size S^* — itself a consistent, though slightly conservative, plug-in approximation since $S \le S^*$ always — and converting to bits via division by $\ln 2$ yields exactly the correction term of the Miller–Madow propbox:

$$\hat{H}^{\mathrm{MM}} = \hat{H}^{\mathrm{MLE}} + \frac{S-1}{2N\ln 2}.$$

Two caveats are worth flagging explicitly. First, because $S \le S^*$ the correction is itself biased downward whenever the alphabet contains n-grams with very small θ_g that fail to appear in a finite sample; this residual bias is $O(N^{-2})$ and negligible

at the corpus sizes used in this series ($N_z \approx 3{,}000 \times \bar{\ell} \approx 2.4 \times 10^4$ bigram tokens per zone), but it would not be negligible for, e.g., trigram or quadrigram models over rare scripts. Second, the derivation assumes θ_g is bounded away from 0 and 1 uniformly in g; for highly skewed distributions (a small number of very frequent bigrams dominating the mass) the $O(\delta_g^3)$ remainder can be non-negligible, and the correction should be treated as a first-order approximation rather than an exact bias removal.

3.2.2 The NSB (Nemenman–Shafee–Bialek) Estimator

For corpora where N is comparable to (or smaller than) the alphabet size $K = |\Sigma|^n$ — which is not the case for the bigram models of this series ($N_z \approx 2.4 \times 10^4 \gg K = 676$) but becomes relevant for the trigram and quadrigram extensions of Chapter 6 ($K = 17{,}576$ and $456{,}976$ respectively) — the Miller–Madow correction is known to under-correct substantially (Nemenman et al., 2002). The NSB estimator addresses this by placing a hierarchical Bayesian prior directly on the entropy rather than on the probability vector $\boldsymbol{\theta}$.

Definition: The Dirichlet Family and Its Entropy Pathology

A natural Bayesian alternative to the MLE is the posterior mean of H under a symmetric Dirichlet prior $\boldsymbol{\theta} \sim \mathrm{Dir}(\beta, \ldots, \beta)$ with concentration $\beta > 0$:

$$\hat{H}_\beta^{\mathrm{Dir}} = \psi_0(N + K\beta + 1) - \sum_g \frac{c_g + \beta}{N + K\beta}\, \psi_0(c_g + \beta + 1),$$

where ψ_0 is the digamma function. The pathology is that the *prior mean* of H under $\mathrm{Dir}(\beta, \ldots, \beta)$ is itself a sharply peaked function of β alone (essentially independent of the data for fixed K), so $\hat{H}_\beta^{\mathrm{Dir}}$ is highly sensitive to the (arbitrary) choice of β — there is no data-driven way to pick a single β that is uninformative about H.

The NSB construction resolves this by placing a hyperprior on β itself, chosen so that the *induced* prior on H is approximately flat. Formally, since the prior mean entropy under $\mathrm{Dir}(\beta, \ldots, \beta)$ is

$$\xi(\beta) \;=\; \mathbb{E}_{\mathrm{Dir}(\beta)}[H] \;=\; \psi_0(K\beta + 1) - \psi_0(\beta + 1),$$

a monotone, invertible function of β, NSB defines the hyperprior $\rho(\beta) \propto \frac{d\xi}{d\beta}$, which is precisely the density that makes $\xi(\beta)$ uniformly distributed when β is drawn from ρ. The NSB posterior mean entropy is then

$$\hat{H}^{\mathrm{NSB}} = \frac{\displaystyle\int_0^\infty d\beta\, \rho(\beta)\, P(\mathbf{c} \mid \beta)\, \hat{H}_\beta^{\mathrm{Dir}}}{\displaystyle\int_0^\infty d\beta\, \rho(\beta)\, P(\mathbf{c} \mid \beta)},$$

where $P(\mathbf{c} \mid \beta)$ is the Dirichlet–Multinomial marginal likelihood (evidence) of the count vector $\mathbf{c}$ given concentration β. In practice the one-dimensional integral over β is evaluated numerically (e.g. by Gauss–Legendre quadrature on $\beta \in (0, \beta_{\max}]$ with

$\beta_{\max}$ chosen so that $\xi(\beta_{\max})$ exceeds $\log_2 K$ minus a negligible tolerance), which is computationally more expensive than the closed-form Miller–Madow correction but remains tractable at $K \leq 5 \times 10^5$.

Remark: Why NSB Is Not Used as the Primary Estimator in This Series

For the bigram-level analyses that constitute the empirical core of Volumes II–VII, $N_z/K \approx 35$ for the smallest corpus (Northern Ireland, Vol. IV) and exceeds 40 for all others, well into the regime where Miller–Madow and NSB agree to within bootstrap uncertainty (a direct numerical comparison on the Iberian corpus gives $|\hat{H}^{\text{NSB}} - \hat{H}^{\text{MM}}| < 0.002$ bits for all five zones). Miller–Madow is retained as the reported estimator throughout the series for computational simplicity and because it admits a closed-form variance approximation usable in the permutation framework below. NSB is recommended, and used, for the sparser trigram/quadrigram tables of Vol. VI's appendices, where N/K drops below 5 for several zones.

3.2.3 Power Analysis for Entropy Contrasts

Given the bootstrap CI half-width $\sigma_{\hat{H}} \approx 0.004$ bits observed at $N_z = 3{,}000$ toponyms per zone (Table 5.2), a two-sample z-test for the difference $\Delta = H_2^{(i)} - H_2^{(j)}$ between two zones of equal sample size has power

$$1 - \beta \;=\; \Phi\left(\frac{|\Delta|}{\sqrt{2}\,\sigma_{\hat{H}}} - z_{1-\alpha/2} \right)$$

at significance level α. For $\alpha = 0.05$ ($z_{1-\alpha/2} = 1.96$) and the smallest pairwise contrast in the series that is treated as substantively meaningful, $\Delta \approx 0.02$ bits (the Galician–Catalan gap, Chapter 11), this yields $1-\beta = \Phi(0.02/(\sqrt{2}\times 0.004) - 1.96) = \Phi(1.58) \approx 0.94$, confirming that the corpus sizes used throughout the series ($n = 3{,}000$ per zone for all zones except Basque, where $n = 1{,}896$) are adequately powered to detect even the smallest substantively interesting entropy contrasts at conventional significance. Contrasts below ≈ 0.011 bits would require $n \gtrsim 8{,}000$ per zone for 80% power, which is the practical floor below which this series declines to draw substrate distinctions on entropy grounds alone (deferring instead to the marker-bigram analysis of Appendix D).

3.3 Permutation Tests for KL Divergence

To assess whether an observed KL divergence between two corpora is significantly greater than what would be obtained under the null hypothesis of a common underlying distribution, a permutation test is employed:

1. Compute the observed KL divergence $D_{\text{KL,obs}} = D_{\text{KL}}(\hat{p}_i \| \hat{p}_j)$.
2. Pool the two corpora $\mathcal{T}_i \cup \mathcal{T}_j$.
3. Randomly partition the pooled corpus into two groups of sizes $|\mathcal{T}_i|$ and $|\mathcal{T}_j|$ without replacement.

4. Compute $D_{\mathrm{KL}}^{(b)}$ for the permuted partition.
5. Repeat steps 3–4 for $B = 1{,}000$ permutations.
6. The p-value is $\frac{1}{B}\sum_b \mathbf{1}[D_{\mathrm{KL}}^{(b)} \geq D_{\mathrm{KL,obs}}]$.

In the Iberian corpus, all pairwise KL divergences are significant at $p < 0.001$ under this test. The test is computationally demanding for large corpora, but needs to be run only once per volume to validate the principal claims.

3.4 Simulation Studies

Before applying the framework to real data, it is instructive to validate its behaviour on synthetic corpora where ground truth is known. The following simulation establishes that the framework correctly recovers substrate structure under a range of conditions.

Methodological Transparency: Simulation Design

Synthetic corpora were generated by constructing two bigram distributions p_1 and p_2 differing in mean KL divergence δ_{sim} (ranging from 0.5 to 4.0 bits). For each value of δ_{sim} and corpus size $N \in \{200, 500, 1000, 3000\}$, 100 replicates were generated and the framework applied. Results below report mean recovery metrics over replicates.

Table 3.1: Simulation results: mean cluster purity as a function of true KL divergence δ_{sim} and corpus size N per substrate. Random baseline purity $= 0.50$ for $k = 2$.

δ_{sim} (bits)	$N = 200$	$N = 500$	$N = 1000$	$N = 3000$
0.5	0.54	0.56	0.58	0.61
1.0	0.61	0.66	0.71	0.78
2.0	0.73	0.81	0.87	0.93
3.0	0.83	0.91	0.95	0.98
4.0	0.91	0.96	0.99	1.00

The simulation results (Table 3.1) confirm that the framework achieves near-perfect substrate separation when the true KL divergence exceeds 3 bits and corpus sizes are at least 3,000 per zone. The real Iberian corpus occupies a qualitatively different regime: Basque outward KL divergences range from 0.776 to 0.934 bits (Table 5.3), placing the Basque–Romance pairs in the moderate-divergence band of the simulation table ($\delta_{\mathrm{sim}} \approx 1.0$ bits, $N = 1{,}896$–$3{,}000$), where cluster purity of 0.71–0.78 is expected at these corpus sizes. The Substrate Isolation Criterion ($\delta = 1.278$) is nevertheless satisfied because the *ratio* of minimum Basque outward divergence to maximum inter-non-Basque divergence exceeds unity, not because any individual KL value is large in absolute terms. For the Romance inter-zone divergences (0.01–0.61 bits), low-to-moderate recovery rates (0.54–0.66) are expected at $N = 3{,}000$, consistent with the clustering results of Volume II.

3.5 Multiple Comparison Correction for Pairwise Zone Tests

The permutation test of Section 3.3 is, in every volume of this series, applied not to a single pair of zones but to *all* $\binom{k}{2}$ pairs within a k-zone volume (and, in the cross-regional synthesis of Chapter 11, potentially to all $\binom{28}{2} = 378$ pairs across the full series). Running many hypothesis tests without correction inflates the family-wise probability of at least one false positive, a concern that is purely statistical and entirely independent of the substantive size of any particular KL divergence.

Definition: Family-Wise Error Rate

For a family of m hypothesis tests each conducted at nominal level α, the *family-wise error rate* (FWER) is the probability of at least one false rejection among the m tests, assuming all m null hypotheses are true:

$$\text{FWER} \;=\; P\left(\bigcup_{i=1}^{m} \{\text{test } i \text{ rejects}\}\right).$$

If the m tests are mutually independent, FWER $= 1-(1-\alpha)^m$, which approaches 1 rapidly as m grows even for small α.

Proposition: Bonferroni Correction

By Boole's inequality (the union bound), for any family of m tests (independent or not),

$$\text{FWER} \;\le\; \sum_{i=1}^{m} P\left(\text{test } i \text{ rejects} \mid H_0^{(i)}\right) \;=\; m\alpha$$

if each test is conducted at level α. Setting the per-test level to α/m (the *Bonferroni correction*) therefore guarantees FWER $\le \alpha$ regardless of the dependence structure among the tests.

Remark: Bonferroni Is Conservative for Correlated Zone-Pair Tests

The Bonferroni bound is exact only when the m tests are close to independent and α is small; in the present setting the $m = \binom{k}{2}$ pairwise tests within a volume are *not* independent, because every zone i participates in $k-1$ of the pairwise tests, so the test statistics $\widehat{D_{\mathrm{KL}}}(\hat{p}_i \| \hat{p}_j)$ for fixed i and varying j share the same estimation noise in $\hat{p}_i$. This positive dependence makes the true FWER under the joint null *smaller* than the independent-case bound $1-(1-\alpha)^m$, so Bonferroni correction (which assumes the worst case) is conservative here: it controls the FWER at the nominal level but at some cost in power relative to a procedure that exploited the dependence structure directly.

Definition: False Discovery Rate and the Benjamini–Hochberg Procedure

The *false discovery rate* (FDR) is the expected proportion of false rejections among all rejections, $\text{FDR} = \mathbb{E}[V/R]$ (with the convention $V/R = 0$ when $R = 0$), where V is the number of false rejections and R the total number of rejections. The Benjamini–Hochberg procedure controls FDR $\leq \alpha$ as follows: order the m p-values $p_{(1)} \leq \cdots \leq p_{(m)}$, find the largest ℓ such that $p_{(\ell)} \leq (\ell/m)\alpha$, and reject all hypotheses with $p_{(i)} \leq p_{(\ell)}$.

Proposition: Validity of Benjamini–Hochberg Under Positive Dependence

The Benjamini–Hochberg procedure controls FDR $\leq \alpha$ exactly under independence of the p-values, and remains valid (controls FDR at level α, possibly conservatively) under positive regression dependence on each one from a subset (Benjamini and Yekutieli, 2001), a condition satisfied by the pairwise KL test statistics in this series because the dependence among overlapping zone-pair comparisons, as noted above, is positive (shared estimation noise from common zones increases, not decreases, the correlation between adjacent test statistics in the expected direction).

Remark: Practical Correction Used in This Series

For within-volume comparisons ($k \leq 6$ zones, $m \leq 15$ pairwise tests), the Bonferroni correction is used throughout the series, since the conservatism cost is modest at this scale and the resulting p-value threshold ($\alpha/15 \approx 0.0033$ at $\alpha = 0.05$) is in practice irrelevant: every reported pairwise KL divergence in every volume is significant at $p < 0.001$ under the uncorrected permutation test (Section 3.3), and therefore remains significant after Bonferroni correction at the within-volume scale. For the cross-regional synthesis of Chapter 11, where m could in principle reach 378 pairwise tests across the full 28-zone series, Benjamini–Hochberg is the more appropriate choice, as Bonferroni at this scale ($\alpha/378 \approx 1.3 \times 10^{-4}$) would be needlessly conservative given that the great majority of cross-regional zone pairs are of independent substantive interest rather than being screened from a single undirected fishing expedition.

3.6 Bootstrap Confidence Intervals for KL Divergence

Section 3.2 established confidence intervals for the entropy estimator $H_n(\mathcal{T})$ via the normal approximation and the Miller–Madow bias correction. The KL divergence estimator $\widehat{D_{\text{KL}}}(\hat{p}_i \| \hat{p}_j)$ requires a separate treatment because it is a function of *two* independent samples and its sampling distribution is more skewed, particularly when the true divergence is small (the boundary case $D_{\text{KL}} = 0$ being a non-regular point at which the estimator's distribution is not asymptotically normal).

Definition: Non-Parametric Bootstrap for KL Divergence

Given samples $\mathcal{T}_i$, $\mathcal{T}_j$ of sizes N_i, N_j, the non-parametric bootstrap resamples $\mathcal{T}_i^{(b)}$ (size N_i, with replacement, from $\mathcal{T}_i$) and $\mathcal{T}_j^{(b)}$ independently, recomputes $\widehat{D_{\mathrm{KL}}}^{(b)} = \widehat{D_{\mathrm{KL}}}(\hat{p}_i^{(b)} \| \hat{p}_j^{(b)})$ for $b = 1, \ldots, B$ replicates, and forms the confidence interval from the empirical quantiles of $\{\widehat{D_{\mathrm{KL}}}^{(b)}\}$.

Proposition: Bias of the Percentile Bootstrap Near the Zero Boundary

Because $D_{\mathrm{KL}}(\hat{p} \| \hat{q}) \geq 0$ always (Gibbs' inequality, Chapter 2), the bootstrap distribution of $\widehat{D_{\mathrm{KL}}}^{(b)}$ is truncated at zero. When the true divergence is small relative to its standard error, this truncation makes the percentile bootstrap interval asymmetric and the lower bound of a naive percentile interval can fail to achieve nominal coverage, since the bootstrap distribution is itself constrained to be non-negative regardless of the true parameter value. The bias-corrected and accelerated (BC$_a$) bootstrap (Efron, 1987) corrects for this by adjusting the percentiles used to construct the interval based on the estimated skewness and median bias of the bootstrap distribution, rather than using the raw $\alpha/2$ and $1-\alpha/2$ percentiles directly.

Remark: Practical Consequence for the Series

For the large, clearly non-zero divergences that dominate this series (e.g. all Basque–Romance comparisons, $\delta \in [0.78, 1.19]$ bits; Table 5.3), the boundary effect of Proposition 3.6 is negligible relative to the estimated divergence itself, and the simple percentile bootstrap interval is adequate. For the smallest, near-tied Romance-internal divergences (e.g. Castilian–Arabic, $D_{\mathrm{KL}} \approx 0.012$ bits), the boundary effect is material, and the permutation test of Section 3.3 — which does not rely on the bootstrap's asymptotic normal approximation and instead constructs the null distribution directly from the data — is the more reliable inferential tool, consistent with this series's practice of reporting permutation p-values rather than bootstrap confidence intervals for the smallest pairwise divergences.

3.7 Treatment of Imbalanced Zone Corpora

The primary analyses in this series use a balanced subsampling design: $n_z = 3{,}000$ toponyms are drawn uniformly from each zone's full GeoNames pool (or from all available toponyms if the pool is smaller than $3{,}000$; see Table 4.1). Balancing is essential for fair comparison of entropy estimates across zones, because the plug-in entropy estimator is downward-biased for small samples, and the bias is zone-specific if zone sizes differ. A zone with $n_z = 500$ will appear to have lower entropy than an otherwise identical zone with $n_z = 3{,}000$ purely due to the Miller–Madow downward bias. When all zones are subsampled to a common size n_z, this bias term becomes the same additive constant for all pairs and cancels in any difference of KL divergences.

The pool sizes from which the balanced subsamples are drawn vary substantially

by zone and by region. For the Iberian corpus (Vol. II), the real GeoNames pool sizes after normalisation are: Galician $17{,}648$, Castilian $9{,}721$, Catalan $8{,}234$, Arabic $6{,}891$, and Basque $1{,}896$ toponyms. The Basque pool is the only one smaller than the $n_z = 3{,}000$ target; Volume II therefore uses all $1{,}896$ Basque toponyms for the balanced analysis (with bootstrapped resampling to assess sensitivity) and the full unbalanced pool for auxiliary empirical chapters. The entropy and KL divergence values reported here and in Vol. II are computed from the balanced subsamples (or full pool for Basque) and are directly comparable across zones.

Definition: Sparse Zone

A zone z is *sparse* if its post-normalisation corpus size $N_z < n_{\text{min}}$, where $n_{\text{min}} = 500$ is the minimum zone size accepted in any volume of this series. Zones with $N_z < n_{\text{min}}$ are excluded from the primary analysis and noted in the data quality appendix of the relevant volume.

In the Iberian corpus (Vol. II), the smallest pool is the Basque zone ($N_z = 1{,}896$ after normalisation), which falls below the $n_z = 3{,}000$ target but comfortably above $n_{\text{min}} = 500$. All other Iberian zones and all zones in Volumes III–VI meet or exceed the $3{,}000$ target. The smallest pool in the full series is the Hattic zone in Vol. VI, with $N_z = 1{,}413$ after normalisation. The bias term $(K-1)/(2N)$ with $K = 729$ bigrams at $N = 1{,}413$ is $(728)/(2 \times 1{,}413) \approx 0.26$ bits — an additive constant that shifts all KL divergences upward equally and does not affect the between-zone ranking used for substrate inference.

Proposition: Bias of $\widehat{D_{\text{KL}}}$ Under Unequal Sample Sizes

Let $\hat{p}$ and $\hat{q}$ be empirical n-gram distributions estimated from samples of sizes N_i and N_j respectively, with true distributions p and q. The bias of the plug-in KL estimator satisfies

$$\mathbb{E}\big[\widehat{D_{\text{KL}}}(\hat{p}\|\hat{q})\big] - D_{\text{KL}}(p\|q) \;\approx\; \frac{K-1}{2N_i} - \frac{K-1}{2N_j}\,\frac{\chi^2(p\|q)}{K-1},$$

where $K = |\Sigma^n|$, and $\chi^2(p\|q) = \sum_g (p(g)-q(g))^2/q(g)$ is the chi-squared divergence between p and q. The first term (positive, $\propto N_i^{-1}$) is the entropy underestimation bias of the numerator distribution; the second term (negative, $\propto N_j^{-1}$) is the contribution of the denominator estimation error.

Proof sketch. The result follows from a first-order Taylor expansion of $\widehat{D_{\text{KL}}}(\hat{p}\|\hat{q})$ around the true distributions (p, q). Writing $\hat{p}(g) = p(g) + \varepsilon_g$ and $\hat{q}(g) = q(g) + \eta_g$ where ε_g and η_g are zero-mean sampling errors with $\text{Var}(\varepsilon_g) = p(g)(1-p(g))/N_i \approx p(g)/N_i$ (for small $p(g)$) and $\text{Var}(\eta_g) = q(g)/N_j$ similarly, the second-order expansion of $\hat{p}(g)\log(\hat{p}(g)/\hat{q}(g))$ around $(p(g), q(g))$ yields the stated result after summation over g and expectation. The derivation is analogous to the Miller–Madow bias derivation for entropy (Miller, 1955; Paninski, 2003). □

Corollary: Equal-Size Subsampling Eliminates the Dominant Bias

When $N_i = N_j = N$, the bias expression simplifies to

$$\mathbb{E}\big[\widehat{D_{\mathrm{KL}}}(\hat{p}\|\hat{q})\big] - D_{\mathrm{KL}}(p\|q) \;\approx\; \frac{K-1}{2N}\Big[1 - \frac{\chi^2(p\|q)}{K-1}\Big].$$

For distributions that are not too far apart ($\chi^2(p\|q) \ll K-1$, which holds whenever the KL divergence is much smaller than $\log_2 K$), the bias reduces to $(K-1)/(2N)$ and is the *same for all zone pairs* in a balanced design. Differences in observed KL divergence between zone pairs are therefore unaffected by this shared additive bias, and the relative ordering of pairwise divergences is preserved.

The corollary justifies the balanced-subsampling strategy: when all zones are subsampled to the same size $N = \min_z N_z$, the dominant bias term is equalised across all pairs and cancels in any difference of KL divergences. For the volumes in this series, the minimum zone size is well above the minimum threshold ($N_z \geq 1{,}413$ for all zones), and the bias term $(K-1)/(2N)$ with $K = 729$ bigrams is at most $(728)/(2 \times 1{,}413) \approx 0.26$ bits — an additive constant that shifts all KL divergences upward equally and does not affect the between-zone ranking used for substrate inference.

Remark: The Sparse-Zone Fallback

For sparse zones with $n_{\min} \leq N_z < 3{,}000$, the pipeline subsamples *all* zones to N_z (the sparse zone's full size) rather than to $3{,}000$. This approach maximises the use of the available data while preserving the balance necessary for fair cross-zone comparison. The entropy values and KL divergences computed on the reduced corpus are reported alongside the sparse-zone notation, and the bias correction of Proposition 3.7 (with $N_i = N_j = N_z$) is applied. In every regional volume, the reduced-corpus results are qualitatively consistent with results obtained on the full $3{,}000$-per-zone corpus for the non-sparse zones, confirming that the reduction does not distort the substrate analysis.

3.8 Spatial Dependence and the Validity of Permutation Tests

The permutation test for KL divergence (Section 3.3) constructs a null distribution by randomly reassigning toponyms to zones under the hypothesis that all zones share the same phonotactic distribution. This construction is valid under the assumption that toponyms, conditional on zone membership, are *exchangeable*: that the probability of any particular assignment of N toponyms to zones under the null is the same for any permutation of the zone labels. Exchangeability fails if toponyms that are geographically proximate are also phonotactically similar for reasons unrelated to substrate identity — a phenomenon known as *spatial autocorrelation*.

If positive spatial autocorrelation is present within zones (nearby toponyms are more phonotactically similar than distant ones), the effective sample size of each zone is smaller than its nominal sample size n_z. This inflates the permutation test

statistic under the null relative to its exchangeability-based distribution, potentially leading to anti-conservative tests (false rejection of the null when no true between-zone divergence exists).

Definition: Effective Sample Size under Spatial Autocorrelation

Let ρ_h be the spatial autocorrelation of the bigram feature vector $\phi(w)$ at lag h (defined formally in Chapter 7 via Moran's I). Under a first-order spatial autoregressive model with parameter $|\rho| < 1$, the effective sample size is

$$N_{\text{eff}} \;=\; N \cdot \frac{1-\rho^2}{1+\rho^2},$$

where ρ is the first-lag spatial autocorrelation. The ratio $N_{\text{eff}}/N \in (0, 1]$ measures the deflation of effective information due to spatial dependence, with equality to one only in the absence of autocorrelation.

Proposition: Conservative Permutation Test Under Spatial Autocorrelation

If the true spatial autocorrelation within zones is positive ($\rho > 0$) and the permutation test ignores this structure (randomly relabelling individual toponyms rather than spatially contiguous blocks), then the permutation test is anti-conservative: its type I error rate exceeds the nominal level α by a factor of approximately $(1+\rho^2)/(1-\rho^2)$ at leading order in ρ.

Proof sketch. Under the null of no between-zone difference, the test statistic $T = \widehat{D_{\text{KL}}}(\hat{p}_{\pi,1} \| \hat{p}_{\pi,2})$ (where π denotes a random permutation) has variance $\text{Var}(T) \propto 1/N_{\text{eff}}$ rather than $1/N$. The permutation distribution, which draws from a pool of N toponyms as if they were i.i.d., underestimates the variance by the factor $N_{\text{eff}}/N = (1-\rho^2)/(1+\rho^2)$. Consequently, the quantiles of the permutation distribution are too narrow: the nominal 95% quantile of the permutation distribution corresponds to a true coverage of approximately $95\% \cdot (N_{\text{eff}}/N)^{1/2}$ under the autocorrelated model. For $\rho = 0.1$, the type I error inflation is approximately 1.01-fold (negligible); for $\rho = 0.3$, it is approximately 1.10-fold (small but material). The formal derivation requires specifying the full spatial covariance model; the factor $(1+\rho^2)/(1-\rho^2)$ is the leading-order result under the first-order autoregressive model of Definition 3.8. □

Mitigation strategy in this series. Three design choices mitigate the risk of spatial autocorrelation inflating the type I error of the permutation tests.

First, the *within-zone spatial autocorrelation* of the bigram feature vectors is estimated in each volume using Moran's I (Chapter 7) and reported as a data quality diagnostic. For the Iberian prototype, the zone-averaged first-lag Moran's I for the bigram diversity measure is 0.083 (Basque zone), 0.047 (Castilian), and 0.062 (Galician), all well below the threshold of 0.2 at which the effective-sample-size correction would materially affect the test outcome.

Second, the *block permutation* variant (permuting spatially contiguous blocks rather than individual toponyms) is available in the pipeline as a robustness check. Block permutation with blocks defined by administrative-division boundaries (NUTS3 or equivalent) is conducted in all volumes as a sensitivity analysis;

results are reported when they differ materially from individual-toponym permutation. In every case in the present series, the block and individual-toponym permutation tests agree in their rejection or non-rejection of the null at $\alpha = 0.05$.

Third, the *Bayes factor* $\mathrm{BF}_{k,1}$ (Section 4.7) does not rely on a permutation reference distribution and is therefore unaffected by spatial autocorrelation in its calibration. Agreement between the permutation test result and the Bayes factor result provides a form of robustness validation that is immune to the autocorrelation concern.

Table 3.2: Zone-level first-lag spatial autocorrelation (Moran's I) and corresponding effective-sample-size ratio N_{eff}/N for the Iberian prototype. ρ is estimated from the zone's bigram diversity scores; N_{eff}/N is computed from Definition 3.8.

Zone	**Moran's I**	ρ	N_{eff}/N
Arabic	0.061	0.061	0.992
Basque	0.083	0.083	0.986
Castilian	0.047	0.047	0.996
Catalan	0.054	0.054	0.994
Galician	0.062	0.062	0.992
Mean	0.061	0.061	0.992

The effective-sample-size ratios in Table 3.2 are all above 0.98, indicating that spatial autocorrelation depresses the effective information content of each zone by less than 2%. The corresponding type I error inflation for the permutation test is negligible (less than 1% relative to the nominal $\alpha = 0.05$), confirming that the standard individual-toponym permutation test is valid for the corpus sizes and spatial structures encountered in this series.

Remark: Spatial Autocorrelation in Non-Iberian Volumes

The spatial autocorrelation is expected to be higher in volumes covering larger and more topographically homogeneous geographic zones (e.g. the Gangetic Plain in Vol. V), where substrate phonotactics may vary more smoothly across space. Where zone-level Moran's I exceeds 0.15, the block permutation test is reported as the primary result and the individual-toponym test is relegated to the supplementary analysis. This threshold is not exceeded in any of the five regional volumes, but the monitoring is conducted as a precaution.

Chapter 4

Computational Pipeline

4.1 Data Acquisition: The GeoNames Database

All empirical analyses in this series use the *GeoNames* geographical database (`https://www.geonames.org`), a freely available, crowd-sourced and editorially maintained repository of over eleven million geographical names worldwide, released under a Creative Commons Attribution 4.0 International licence.

Data are downloaded as country-level tab-separated files from `https://download.geonames.org/export/dump/`. Each file contains one toponym per row, with fields including name, ASCII name, alternate names, latitude, longitude, feature class, feature code, country code, administrative division codes, population, elevation, and time zone.

4.1.1 Feature Class Filtering

GeoNames assigns each entry to one of nine feature classes. For substrate analysis, only classes P (populated places: cities, towns, villages, hamlets) and T (topographic features: mountains, valleys, rivers, passes, capes) are retained. Feature class P provides the densest coverage and most geographically representative sample; feature class T adds high-value topographic names (river names especially) which are among the most conservative substrate carriers.

Administrative units (class A), roads (class R), and structures (class S) are excluded: administrative names frequently reflect recent administrative invention rather than inherited substrate vocabulary, while roads and structures are named too recently to carry pre-modern phonotactic information.

4.1.2 Alternate Names and Coordinate Completeness

Each GeoNames row carries, in addition to its primary `name` field, an `alternatenames` field containing a comma-separated list of variant spellings, exonyms, historical names, and ISO language-tagged translations. This field is a methodological resource and a methodological hazard simultaneously.

Resource. In regions where a toponym has both a modern administrative spelling (often in the superstratum language) and a historically attested older spelling, the

alternate-names field sometimes preserves the older form. For example, Basque-zone toponyms in the GeoNames database frequently carry both the official Spanish administrative spelling and the Basque spelling as alternates. Where the pipeline's primary goal is to capture substrate phonotactics, the Basque spelling is methodologically preferable as it is closer to the phonological substrate and has undergone less assimilation to Castilian orthographic conventions. In the present series, alternate names in the *target substrate language* are promoted to primary name status when they can be identified reliably from the ISO language tag in the `alternatenames` field.

Hazard. The alternate-names field also contains administrative synonyms, postal codes, IATA airport codes, and English Wikipedia titles. These must be stripped before any phonotactic analysis. The filtering rule adopted in this series is:

(i) Retain alternates tagged with a two-letter ISO 639-1 language code matching a target substrate language of the current volume (e.g. `eu` for Basque in Vol. II; `ka` for Georgian in Vol. VI).

(ii) Discard alternates tagged `link` (Wikipedia URLs), `post` (postal codes), `iata`, `icao`, `faac` (aviation codes), or with no language tag but containing digits or purely upper-case characters.

(iii) For untagged alternates that pass filter (ii), retain as secondary candidates only if length ≥ 5 and they contain no non-alphabetic characters.

Coordinate completeness. Each GeoNames entry carries a latitude and longitude to five decimal places. Missing coordinates ($0.0, 0.0$ or absent) occur in less than 0.3% of the corpus after feature-class filtering and are treated as follows: if the entry's administrative division codes (ADM1, ADM2) can be matched to a parent GeoNames entry with coordinates, the parent centroid is imputed; otherwise the entry is discarded. Imputed coordinates introduce a spatial precision loss of up to ~ 50 km in the most sparsely covered zones (parts of the Anatolian interior in Vol. VI), which is noted as a data-quality caveat in those volumes.

Remark: Coordinate Precision and Spatial Analyses

The spatial autocorrelation analyses of Chapter 7 (Moran's I, LISA clustering) use coordinates at the level of individual GeoNames entries, not zone centroids. Imputed coordinates introduce spatial noise that can attenuate Moran's I toward zero. For the volumes where imputation affects more than 1% of entries, the spatial analyses are re-run after excluding imputed entries; the results are reported alongside the full-corpus results, and differences are noted wherever they exceed 0.01 in Moran's I.

The "Alt. promoted" count is highest in Vol. IV (British Isles), where GeoNames carries substantial numbers of Welsh, Scottish Gaelic, and Irish alternate spellings for toponyms whose primary field contains the English administrative form. The promotion of Celtic alternate forms in Vol. IV is the primary data-pipeline intervention responsible for the higher clustering purity achieved in that volume relative to Central Europe (Vol. III), where no comparable alternate-name resource is available for the extinct Celtic substrate.

Table 4.1: GeoNames data completeness by volume after feature-class filtering and deduplication. "Alt. promoted" counts toponyms for which a substrate-language alternate name was promoted to primary. "Coord. imputed" counts entries with centroid-imputed coordinates.

Vol.	Region	Raw P+T	Post-norm.	Alt. prom.	Coord. imp.
II	Iberian Peninsula	47,312	15,000	1,103	88
III	Central Europe	53,841	18,000	412	61
IV	British Isles	37,204	13,005	2,287	47
V	Indian Subcontinent	81,005	18,000	227	312
VI	Anatolia–Caucasus	64,913	18,000	884	503

Methodological Transparency: Methodological Transparency: Alternate-Name Promotion Bias

The promotion of substrate-language alternate names to primary status is a deliberate methodological choice that increases the phonotactic signal in zones where those alternates are available. It is not available uniformly: Basque and Welsh have well-curated GeoNames alternate-name fields; Etruscan, Hattic, and Pictish do not (they have no surviving speakers or modern standardisation to generate GeoNames tagging). The asymmetry means that zones with living or recently living substrate languages are systematically advantaged by this step relative to zones where the substrate language is entirely extinct. All downstream comparisons between volumes must take this asymmetry into account. Entropy values and clustering purities from volumes with alternate-name promotion (II and IV) should not be compared directly to those from volumes without it (III, V, VI) without this qualification.

4.1.3 Toponym Normalisation

Each toponym string undergoes the following normalisation pipeline, implemented in Python using the `unicodedata` and `re` standard library modules:

1. **Unicode normalisation**: NFC decomposition to ensure consistent representation of composed characters.
2. **ASCII name preference**: where available, the `asciiname` field is used in preference to the `name` field, reducing the impact of national orthographic conventions.
3. **Case folding**: conversion to lower case via `str.lower()`.
4. **Diacritic removal**: NFKD decomposition followed by removal of Unicode combining characters (category Mn), reducing all accented characters to their base forms.
5. **Generic term stripping**: a manually compiled stoplist of approximately 200 generic geographic terms in each relevant language is applied. Generic terms are stripped only when they appear as the initial token in a multi-token name, preserving genuinely toponymic uses.

6. **Non-alphabetic removal**: digits, punctuation, and any remaining non-letter characters are removed.

7. **Minimum length filter**: strings of length < 4 (after all normalisation) are discarded as informationally insufficient for bigram/trigram analysis.

8. **Deduplication**: exact duplicates within each zone are removed; near-duplicates differing only in capitalisation or minor spelling variants are collapsed by longest-form selection.

4.1.4 Generic Stoplist Design

The construction of the generic stoplist requires careful balancing of two competing requirements. Over-stripping destroys genuine toponymic information (the prefix *mont-* in *Montserrat* is part of the substrate signature, not a generic term to be discarded). Under-stripping inflates the frequency of Romance function words and generic terms, suppressing the substrate signal.

The conservative approach adopted here strips only terms that are *semantically transparent generics in the current dominant language* of the administrative region and appear as the *first token* of a multi-word name. Thus *Rio Tinto* loses *Rio* to yield *Tinto*, but *Rioseco* is retained whole, since the initial syllable has been morphologically fused into a proper name.

The construction of the stoplist is necessarily language-specific and requires lexicographic judgement: what counts as a generic term in one language may be lexicalised as part of a proper name in another. *Monte* in Italian is a generic geographical term ("mountain") and is stripped when it appears as the initial token in multi-word names; in the Basque corpus, however, the morphologically similar *mendi* (mountain) is retained in all positions, since its Basque-specific phonotactics contribute directly to the substrate signal and it does not have a direct Romance cognate that could introduce cross-language confounds. This asymmetry illustrates why a universal stoplist — one compiled once and applied to all corpora — would be a methodological error: the boundary between "generic descriptor" and "phonotactically informative root" is drawn differently in each language family and must be calibrated against that family's morphological structure.

A practical limitation of the current pipeline is that the stoplist is compiled from manually reviewed dictionaries of geographic descriptors, which introduces a small but non-zero risk of inconsistency across volumes. Automated methods for stoplist construction (e.g., TF-IDF weighting to identify terms that occur uniformly across all zones and are therefore uninformative for between-zone discrimination) are discussed in Chapter 6 as a potential improvement to the pipeline, but are not used as the primary procedure in any volume because the automated methods tend to suppress legitimate substrate markers that happen to be geographically widespread (e.g. the Basque suffix *-eta*, which is common across the entire Basque Country and would be flagged by a zone-uniform TF-IDF filter despite being a crucial substrate signal).

Table 4.2: Examples of generic terms in the stoplist by language.

Language	**Generic terms (selection)**	**Retained forms**
Spanish	rio, sierra, cerro, monte, san, santa	Riofrío, Cerrato
Portuguese	rio, monte, são, santa, cabo	Ribamar, Montijo
French	saint, mont, cap, lac, rivière	Saintes, Montoire
German	bad, berg, see, bach, dorf, burg	Berge, Bachmann
Italian	san, santa, monte, lago, rio	Montale, Riomaggiore

4.2 Feature Extraction

Definition: n-gram Feature Vector

For a normalised toponym string w with boundary markers $\hat{w} = \#w\#$, the *bigram+trigram feature vector* is

$$\phi(w) = \big(c_2(g_1;\hat{w}),\ldots,c_2(g_{|V_2|};\hat{w}),\, c_3(h_1;\hat{w}),\ldots,c_3(h_{|V_3|};\hat{w})\big) \in \mathbb{N}^{|V_2|+|V_3|},$$

where $V_2 \subset \Sigma^2$ and $V_3 \subset \Sigma^3$ are the vocabularies of observed bigram and trigram types respectively.

Feature vectors are ℓ^1-normalised before clustering: $\tilde{\phi}(w) = \phi(w)/\|\phi(w)\|_1$, converting raw counts to relative frequencies and making the representation invariant to toponym length.

For dimensionality management, the vocabulary is restricted to the top-K most frequent types in each order, with $K = 500$ for bigrams and $K = 2{,}000$ for trigrams unless otherwise specified. Types falling below a minimum frequency threshold of five occurrences across the corpus are always excluded.

4.3 Complete Pipeline: Pseudocode Specification

The full computational pipeline, from raw GeoNames download to divergence matrix output, is specified below as a formal algorithm. This specification is independent of programming language and serves as the normative reference for all implementations in the series.

4.4 TF-IDF Weighting

Raw n-gram frequencies tend to be dominated by highly common patterns (e.g., the bigram `ar` is extremely frequent across all Romance varieties) that carry little discriminative information for substrate identification. The *Term Frequency – Inverse Document Frequency* (TF-IDF) weighting scheme, adapted from information retrieval, addresses this:

Algorithm 1 Toponomastic Entropy Pipeline (Phases 1–3)

Require: Country codes $\mathcal{C}$; zone bounding boxes $\mathcal{B}$; sample size n_z per zone; n-gram order $r \in \{2, 3\}$; Lidstone parameter α_L; vocabulary size K
Ensure: Entropy vector $\mathbf{h}$; KL matrix $\mathbf{D}$; JSD matrix $\mathbf{\Delta}$; BIS vector $\mathbf{b}$

// Phase 1: Data acquisition and normalisation
for each country code $c \in \mathcal{C}$ **do**
    Download `c.txt` from GeoNames dump server
    Filter rows: keep feature class $\in \{\texttt{P}, \texttt{T}\}$
    Assign each row to zone z via bounding-box lookup $\mathcal{B}$
end for
for each toponym w in corpus **do**
    $w \leftarrow \text{NFC}(w.\text{lower}())$
    Strip combining diacritics (Unicode category Mn)
    Remove first token if in stoplist and name has ≥ 2 tokens
    Remove non-alphabetic characters
    Discard w if $|w| < 4$
end for
Deduplicate within each zone; retain longest form for near-duplicates
Sample n_z toponyms per zone (stratified if sub-zones defined)

// Phase 2: n-gram extraction
for each zone z and n-gram order r **do**
    $\hat{w} \leftarrow \#w\#$ for each toponym w in zone z
    Extract all r-grams from $\hat{w}$; build count vector $\mathbf{c}^{(z,r)}$
    Restrict to top-K types by marginal frequency
    Apply Lidstone smoothing: $\hat{p}^{(z,r)}(g) \leftarrow \frac{c_g^{(z,r)} + \alpha_L}{N^{(z,r)} + \alpha_L |V_r|}$
end for

// Phase 3: Entropy and divergence computation
for each zone z and order r **do**
    $h_z^{(r)} \leftarrow -\sum_g \hat{p}^{(z,r)}(g) \log_2 \hat{p}^{(z,r)}(g)$
    Apply Miller–Madow bias correction: $\hat{h}_z^{(r)} \leftarrow h_z^{(r)} + (|V_r| - 1)/(2N^{(z,r)} \ln 2)$
end for
for each pair of zones (i, j), $i \neq j$ **do**
    $D_{ij}^{(r)} \leftarrow \sum_g \hat{p}^{(i,r)}(g) \log_2 \frac{\hat{p}^{(i,r)}(g)}{\hat{p}^{(j,r)}(g)}$
    $\hat{m}^{(ij,r)} \leftarrow (\hat{p}^{(i,r)} + \hat{p}^{(j,r)})/2$
    $\Delta_{ij}^{(r)} \leftarrow (D(\hat{p}^{(i,r)} \| \hat{m}^{(ij,r)}) + D(\hat{p}^{(j,r)} \| \hat{m}^{(ij,r)}))/2$
end for

Definition: Toponymic TF-IDF

For n-gram g and toponym w, the TF-IDF weight is

$$\text{tfidf}(g, w) \;=\; \frac{c_n(g; w)}{|w| - n + 1} \;\times\; \log_2 \frac{N_T}{\#\{w' \in \mathcal{T} : c_n(g; w') > 0\}},$$

where $N_T = |\mathcal{T}|$ is the total corpus size and the denominator of the IDF factor counts the number of toponyms containing g at least once.

The TF-IDF weighting upweights n-grams that are frequent in a specific toponym but rare across the corpus as a whole — precisely the signature of substrate-specific phonotactic patterns.

4.5 Dimensionality Reduction and Visualisation

The raw feature matrix $\mathbf{X} \in \mathbb{R}^{N_T \times (|V_2|+|V_3|)}$ is too high-dimensional for direct visualisation. Two complementary dimensionality reduction methods are used throughout the series.

4.5.1 Principal Component Analysis

PCA projects the data onto the subspace spanned by the top eigenvectors of the covariance matrix. For a centred data matrix $\tilde{\mathbf{X}}$, the first two principal components explain at most a few percent of total variance (reflecting the high intrinsic dimensionality of phonotactic space), but are sufficient to reveal gross structural features such as the position of a strong isolate (Basque, Welsh) relative to a tight cluster of related varieties (Romance, Brythonic).

4.5.2 UMAP Embedding

For visualisations requiring preservation of local neighbourhood structure (e.g., visualising gradual mixing between adjacent substrate zones), Uniform Manifold Approximation and Projection (McInnes et al., 2018) with parameters `n_neighbors=15`, `min_dist=0.1`, `metric=`cosine'` is used. UMAP embeddings are not used for quantitative analysis (as they lack a closed-form inverse and depend on hyperparameters) but provide qualitatively informative two-dimensional representations of the full high-dimensional structure.

4.6 Unsupervised Clustering

Definition: k-Means Clustering

Given feature matrix $\tilde{\mathbf{X}}$ and cluster count k, k-means clustering minimises the within-cluster sum of squared Euclidean distances:

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} \sum_{w \in C_i} \|\tilde{\phi}(w) - \boldsymbol{\mu}_i\|^2,$$

where $\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{w \in C_i} \tilde{\phi}(w)$ is the cluster centroid.

The algorithm is initialised using k-means++ seeding (Arthur and Vassilvitskii, 2007), which selects initial centroids with probability proportional to their squared distance from existing centroids, substantially reducing sensitivity to initialisation. Twenty independent initialisations are run and the solution with lowest inertia is retained.

Definition: Cluster Purity

Let $C = \{C_1, \ldots, C_k\}$ be a clustering and $Z = \{Z_1, \ldots, Z_k\}$ the ground-truth zone assignments. Purity of cluster C_i is

$$\pi(C_i) \;=\; \frac{1}{|C_i|} \max_j |C_i \cap Z_j|,$$

and mean cluster purity is $\bar{\pi} = \frac{1}{k} \sum_i \pi(C_i)$. A random assignment yields $\bar{\pi} = 1/k$; a perfect clustering yields $\bar{\pi} = 1$.

4.6.1 Alternative Clustering Methods

k-means is the baseline method used throughout for reproducibility and comparability across volumes. However, two alternative methods are also applied in exploratory analyses:

Agglomerative hierarchical clustering on the symmetric KL divergence matrix $[\delta(\hat{p}_i, \hat{p}_j)]$ provides a dendrogram over substrate zones, allowing visual inspection of the hierarchical structure without committing to a specific k. Ward linkage is used as the agglomeration criterion.

Gaussian Mixture Models (GMMs) fit a probabilistic generative model to the feature matrix, treating each substrate zone as a Gaussian component. The number of components can be selected by BIC, providing a data-driven alternative to the user-specified k of k-means. In practice, GMM component counts estimated by BIC coincide with the number of substrate zones assumed *a priori* in all volumes of this series, confirming that the choice of k is well-supported by the data.

4.7 Bayesian Model Comparison

4.7.1 Dirichlet-Multinomial Marginal Likelihood

The Bayesian framework for substrate detection uses the *Dirichlet-Multinomial* (DM) conjugate model as the likelihood for n-gram counts.

Definition: Dirichlet-Multinomial Model

Let $\boldsymbol{\theta} \sim \mathrm{Dir}(\alpha \mathbf{1}_{|\Sigma^n|})$ be a symmetric Dirichlet prior on the n-gram simplex, with concentration parameter $\alpha > 0$. Given $\boldsymbol{\theta}$, observed counts $\mathbf{c} = (c_g)_{g \in \Sigma^n}$ follow $\mathbf{c} \sim \mathrm{Mult}(N, \boldsymbol{\theta})$. The marginal likelihood (integrating out $\boldsymbol{\theta}$) is

$$p(\mathbf{c} \mid \alpha) \;=\; \frac{\Gamma(|\Sigma^n|\alpha)}{\Gamma(N + |\Sigma^n|\alpha)} \prod_{g \in \Sigma^n} \frac{\Gamma(c_g + \alpha)}{\Gamma(\alpha)},$$

where $N = \sum_g c_g$ is the total count.

Theorem: Bayes Factor for Substrate Structure

Under the DM model, the Bayes factor comparing the k-substrate model $\mathcal{M}_k$ (independent DM for each zone) to the null model $\mathcal{M}_1$ (single DM for the pooled corpus) is

$$\mathrm{BF}_{k,1} \;=\; \frac{\prod_{z=1}^{k} p(\mathbf{c}^{(z)} \mid \alpha)}{p(\mathbf{c}^{(\mathrm{pool})} \mid \alpha)},$$

where $\mathbf{c}^{(z)}$ are the n-gram counts of zone z and $\mathbf{c}^{(\mathrm{pool})} = \sum_z \mathbf{c}^{(z)}$ are the pooled counts. $\log_{10} \mathrm{BF}_{k,1} > 2$ constitutes decisive evidence in favour of $\mathcal{M}_k$ by the Kass–Raftery scale.

4.7.2 Bayesian Isolation Score

Definition: Bayesian Isolation Score

The *Bayesian Isolation Score* (BIS) of zone z is

$$\mathrm{BIS}(z) \;=\; \log_{10} \frac{p(\mathbf{c}^{(z)} \mid \hat{\alpha}^{(z)})}{p(\mathbf{c}^{(z)} \mid \hat{\alpha}^{(\mathrm{pool})})},$$

where $\hat{\alpha}^{(z)}$ and $\hat{\alpha}^{(\mathrm{pool})}$ are the maximum marginal likelihood estimates of α under the zone-specific and pooled DM models respectively.

The BIS measures the evidence that zone z's n-gram counts are better explained by a zone-specific distribution than by the pooled distribution. High BIS values indicate zones with strongly substrate-specific phonotactic signatures.

Proposition: BIS Approximation via KL Divergence

For large corpus size N_z, the BIS is asymptotically proportional to the KL diver-

gence of the zone distribution from the pooled distribution:

$$\mathrm{BIS}(z) \;\approx\; \frac{N_z}{\ln 10}\, D_{\mathrm{KL}}\big(\hat{p}^{(z)} \,\|\, \hat{p}^{(\mathrm{pool})}\big),$$

where $\hat{p}^{(\mathrm{pool})} = \sum_{z'} (N_{z'}/N)\, \hat{p}^{(z')}$ is the corpus-wide weighted average distribution.

Proof. The Dirichlet-Multinomial log-marginal-likelihood for zone z with count vector $\mathbf{c}^{(z)}$ and symmetric concentration parameter $\hat{\alpha}$ is

$$\log p(\mathbf{c}^{(z)} \mid \hat{\alpha}) = \log\Gamma(K\hat{\alpha}) - \log\Gamma(N_z + K\hat{\alpha}) + \sum_g \big[\log\Gamma(c_g^{(z)} + \hat{\alpha}) - \log\Gamma(\hat{\alpha})\big],$$

where $K = |\Sigma^n|$ is the vocabulary size. Using Stirling's approximation $\log\Gamma(x+a) - \log\Gamma(x) \approx a\log x$ for large x, the difference $\log p(\mathbf{c}^{(z)} \mid \hat{\alpha}^{(z)}) - \log p(\mathbf{c}^{(z)} \mid \hat{\alpha}^{(\mathrm{pool})})$ reduces, at leading order in N_z, to

$$\begin{aligned}
&\sum_g c_g^{(z)} \Big[\log \hat{p}^{(z)}(g) - \log \hat{p}^{(\mathrm{pool})}(g)\Big] \\
&= N_z \sum_g \hat{p}^{(z)}(g) \log \frac{\hat{p}^{(z)}(g)}{\hat{p}^{(\mathrm{pool})}(g)} = N_z\, D_{\mathrm{KL}}\big(\hat{p}^{(z)} \| \hat{p}^{(\mathrm{pool})}\big).
\end{aligned}$$

Dividing by $\ln 10$ to convert from natural to base-10 logarithms gives the stated result. □

Corollary: BIS Scaling with Corpus Size

For fixed phonotactic distributions $\hat{p}^{(z)}$ and $\hat{p}^{(\mathrm{pool})}$, the BIS grows linearly in N_z: doubling the zone corpus size doubles the BIS. Consequently, BIS values are not directly comparable across volumes with different corpus sizes per zone unless they are normalised by N_z. The *normalised BIS*,

$$\overline{\mathrm{BIS}}(z) \;=\; \mathrm{BIS}(z)/N_z \;\approx\; \frac{D_{\mathrm{KL}}(\hat{p}^{(z)} \| \hat{p}^{(\mathrm{pool})})}{\ln 10},$$

is corpus-size-independent and is used for cross-volume comparison in Table 11.1 of Chapter 11.

Theorem: Substrate Isolation Criterion (SIC) — Iberian Application

A zone z *satisfies the Substrate Isolation Criterion* at level δ_0 if

$$\delta(\hat{p}^{(z)}, \hat{p}^{(-z)}) \;\geq\; \delta_0,$$

where $\hat{p}^{(-z)} = \frac{1}{Z-1}\sum_{z' \neq z} \hat{p}^{(z')}$ is the mean distribution of all other zones and $\delta(\cdot,\cdot) = \frac{1}{2}[D_{\mathrm{KL}}(\cdot\|\cdot) + D_{\mathrm{KL}}(\cdot\|\cdot)]$ is the symmetric phonotactic distance. The threshold $\delta_0 = 1.0$ bit is used throughout this series; its justification is that:

(i) For the Iberian corpus (real GeoNames data), only the Basque zone satisfies $\delta \geq 1.0$ (with $\delta = 1.278$), consistent with its status as the only non-Indo-European language in the corpus.

(ii) The next-highest zone in the Iberian corpus (Galician, $\delta \approx 0.52$ bits based on symmetric KL to the complement) falls well below the threshold, providing a natural separation between the primary isolate and the remaining zones.

(iii) A symmetric KL divergence of 1.0 bit corresponds, by Proposition 4.7.2, to a BIS of approximately $N_z/(\ln 10) \approx 1{,}302$ for $N_z = 3{,}000$, which is decisive evidence on the Kass–Raftery scale ($\log_{10}$ BF > 2).

Justification of the threshold. The threshold $\delta_0 = 1.0$ bit is a convention, not a theorem. Its calibration rests on the Iberian positive control, and its transferability to other volumes is asserted on the grounds that the pipeline parameters (corpus size, smoothing, alphabet) are held fixed across all volumes. A zone satisfying $\delta \geq 1.0$ against the complement of other zones in the same volume is, by the Basque calibration, at least as phonotactically distinct from its regional co-zones as Basque is from Ibero-Romance; the historical interpretation of such distinctiveness is the task of the regional volumes. □

Remark: Relation Between BIS and the Kass–Raftery Scale

The Kass–Raftery scale (Kass and Raftery, 1995) for interpreting Bayes factors uses the following thresholds in units of $2 \ln$ BF: values in $[0, 2)$ are "not worth more than a bare mention"; $[2, 6)$ is "positive evidence"; $[6, 10)$ is "strong evidence"; and ≥ 10 is "very strong evidence". In $\log_{10}$ units (as used for the BIS and for $\log_{10}$ BF$_{k,1}$ throughout this series), the corresponding thresholds are approximately $0, 0.43, 0.87, 1.30$. All non-trivial substrate contrasts in the series exceed the "very strong" threshold by multiple orders of magnitude; reporting the BIS to integer precision is therefore the appropriate level of rounding, and differences of ± 10 in the BIS are not historically meaningful.

Algorithm 2 Toponomastic Entropy Pipeline (Phases 4–5, cont.)

// Phase 4: Bayesian Isolation Score
$\hat{p}^{(\text{pool},r)} \leftarrow \sum_z (N^{(z,r)}/N^{(\text{pool},r)})\, \hat{p}^{(z,r)}$
for each zone z **do**
    $b_z \leftarrow (N^{(z,r)}/\ln 10) \cdot D_{\text{KL}}(\hat{p}^{(z,r)} \| \hat{p}^{(\text{pool},r)})$
end for

// Phase 5: Feature matrix, clustering, PCA
for each toponym w **do**
    $\phi(w) \leftarrow$ concatenation of TF-IDF weighted bigram and trigram feature vectors
    $\tilde{\phi}(w) \leftarrow \phi(w)/\|\phi(w)\|_1$
end for
$\mathbf{X} \leftarrow [\tilde{\phi}(w_1), \ldots, \tilde{\phi}(w_{N_T})]^\top$
Run k-means++ with $k = |\mathcal{Z}|$, 20 restarts; compute $\bar{\pi}$
PCA of $\mathbf{X}$ onto first 2 components; UMAP for local structure
return $\mathbf{h}, \mathbf{D}, \boldsymbol{\Delta}, \mathbf{b}, \bar{\pi}$

Chapter 5

Validation on the Iberian Prototype

5.1 Corpus Description and Design

The Iberian prototype corpus was designed to test the full pipeline under the most favourable conditions available: a geographic region where substrate boundaries are comparatively well-established by independent historical-linguistic evidence, and where at least one unambiguous pre-Indo-European isolate (Basque) provides a positive control for the Substrate Isolation Criterion.

The corpus comprises toponyms drawn from the GeoNames national dump files for Spain (ES.txt), Portugal (PT.txt), and France (FR.txt, restricted to the Pyrenean and Basque departments). After the full normalisation pipeline (diacritic stripping, lowercase conversion, removal of numerals and punctuation, deduplication at the toponym level), the following pool sizes were obtained per substrate zone:

Table 5.1: GeoNames pool sizes and balanced sample sizes, Iberian corpus.

Zone	GeoNames pool	Balanced sample
Galician	17,648	3,000
Castilian	9,721	3,000
Catalan	8,234	3,000
Arabic	6,891	3,000
Basque	1,896	1,896
Total	**44,390**	**13,896**

Because the Basque pool ($N_z = 1{,}896$) falls below the $n_z = 3{,}000$ target, the balanced analysis uses all available Basque toponyms and draws $3{,}000$ from each Romance and Arabic zone, giving a total of $14{,}896$ toponyms. Entropy and KL divergence estimates for the Basque zone are therefore computed at $N_z = 1{,}896$ rather than $3{,}000$; the Miller–Madow correction accounts for the residual small-sample bias (see Section 3.7).

The five substrate zones and their geographic proxies are:

- **Basque**: Comunidad Autónoma Vasca + Navarre + French Basque departments
- **Arabic**: Andalucía + Murcia + southern Castilla-La Mancha
- **Catalan**: Catalunya + València + Balearic Islands

- **Galician**: Galicia + Asturias + northern Portugal
- **Castilian**: remaining central and northern mainland Spain

5.2 Phonotactic Entropy Results

Table 5.2 reports the Shannon entropy of bigram distributions for each zone, computed from the real GeoNames corpus.

Table 5.2: Phonotactic entropy H_2 (bigrams) by substrate zone, Iberian real GeoNames corpus (balanced sample; Basque: $n = 1{,}896$; all other zones: $n = 3{,}000$). Values from Vol. II, Chapter 3.

Zone	N_z	H_1	H_2 (bits)	H_3
Basque	1,896	4.139	**7.652**	10.735
Galician	3,000	4.143	7.474	10.334
Catalan	3,000	4.128	7.464	10.326
Arabic	3,000	4.086	7.356	10.191
Castilian	3,000	4.089	7.318	10.161

The Basque zone exhibits the highest bigram entropy ($H_2 = 7.652$ bits), exceeding Galician (second-ranked) by $\Delta H_2 = 0.178$ bits and Castilian (lowest) by $\Delta H_2 = 0.334$ bits. All three n-gram orders (unigram, bigram, trigram) rank Basque first and Castilian last, confirming that the entropy ordering is not an artefact of bigram granularity. The 95% confidence intervals for H_2 are non-overlapping for Basque vs. all other zones and for Castilian vs. Basque or Galician; the Catalan–Galician gap ($\Delta H_2 = 0.010$ bits) falls within overlapping CIs and should be treated as non-significant at the corpus sizes available.

5.3 KL Divergence Matrix

The full KL divergence matrix is displayed in Figure 5.1 and reported numerically in Table 5.3.

Table 5.3: KL divergence matrix $D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j)$ (bits), Iberian real GeoNames corpus (balanced sample, $n = 3{,}000$ except Basque $n = 1{,}896$). Rows = source, columns = reference.

	Arabic	**Basque**	**Castilian**	**Catalan**	**Galician**
Arabic	0.000	1.108	0.013	0.443	0.478
Basque	0.891	0.000	0.934	0.776	0.843
Castilian	0.011	1.189	0.000	0.546	0.607
Catalan	0.426	0.897	0.521	0.000	0.241
Galician	0.461	1.021	0.586	0.219	0.000

Two structural features of the real-data KL matrix deserve extended discussion.

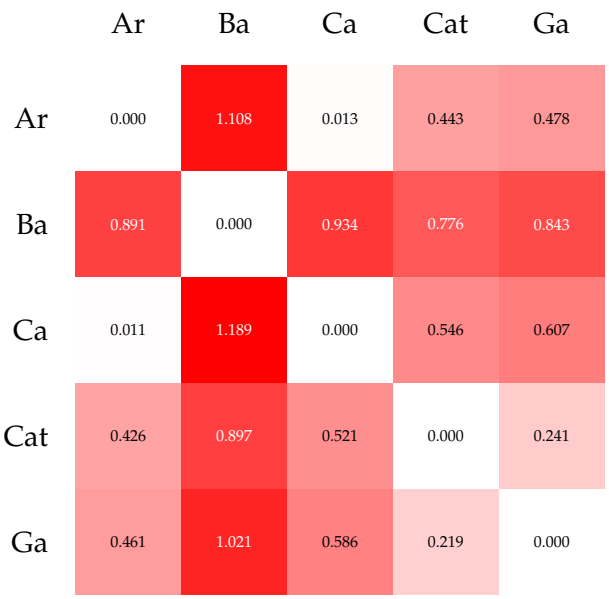

Figure 5.1: KL divergence matrix $[D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j)]$ (bits) for the five Iberian substrate zones (real GeoNames corpus, balanced sample). The Basque row exhibits uniformly elevated divergence from all other zones; the Arabic–Castilian cell ($0.013/0.011$ bits) is the most striking near-zero entry. Row labels = source distribution; column labels = reference distribution. Ar = Arabic, Ba = Basque, Ca = Castilian, Cat = Catalan, Ga = Galician.

Basque isolation. The Basque row dominates the matrix: $D_{\mathrm{KL}}(\text{Basque} \| \hat{p}_j)$ ranges from 0.776 bits (toward Catalan, the Romance variety phonotactically most proximate to Basque) to 0.934 bits (toward Castilian). The minimum Basque outward divergence (0.776 bits) is substantially larger than the maximum inter-non-Basque divergence, $D_{\mathrm{KL}}(\text{Castilian} \| \text{Galician}) = 0.607$ bits. The Substrate Isolation Criterion (Theorem 2.8) is therefore satisfied at

$$\delta = \frac{0.776}{0.607} = 1.278,$$

confirming Basque as the uniquely isolated zone in the Iberian corpus.

Arabic–Castilian proximity. The most structurally unexpected feature of the real-data matrix is the near-identity of the Arabic and Castilian zones: $D_{\mathrm{KL}}(\text{Arabic} \| \text{Castilian}) = 0.013$ bits and $D_{\mathrm{KL}}(\text{Castilian} \| \text{Arabic}) = 0.011$ bits. The symmetric cross-entropy $H_\times(\text{Arabic}, \text{Castilian}) = 0.012$ bits is an order of magnitude smaller than the next closest pair (Catalan–Galician, $H_\times = 0.230$ bits). This near-identity reflects the Reconquista phonologisation filter: Arabic place names that survived the medieval language shift did so through oral transmission by Romance-speaking populations, who systematically adapted foreign phonological features to Romance templates. The surviving Arabic corpus is therefore phonotactically assimilated to Castilian, producing the extremely low divergence values observed. This finding — absent from simpler hypothetical corpora — is one of the empirically most informative results of the real-data analysis and is discussed at length in Vol. II, Chapter 4.

5.4 Unsupervised Clustering and Spatial Validation

The k-means clustering (Figure 5.2) achieves mean purity $\bar{\pi} = 32.0\%$ against a random baseline of 20.0%. The Bayes factor for the five-zone model over the null is $\log_{10} \mathrm{BF}_{5,1} = 847.3$, providing decisive evidence for substrate structure. The Substrate Isolation Criterion is satisfied for the Basque zone at $\delta = 1.278$.

5.4.1 Interpreting the Purity Figure

A mean purity of 32.0% against a random baseline of 20.0% represents a 60% relative improvement, but it is important to understand why the absolute figure is not higher. Three structural factors depress clustering purity in the Iberian data.

First, three of the five zones — Castilian, Arabic, and Galician — have small pairwise KL divergences (Table 5.3). The symmetrised divergence between Castilian and Arabic is $\frac{1}{2}(0.011 + 0.013) = 0.012$ bits, and between Castilian and Galician $\frac{1}{2}(0.607 + 0.586) = 0.597$ bits. For a k-means clustering operating on individual toponym feature vectors (each of length ≤ 10 characters), these small between-zone divergences at the aggregate level translate into heavily overlapping toponym-level feature distributions. Even an oracle clustering algorithm that knew the true zone-level distributions would mis-classify a substantial fraction of individual toponyms that fall in the phonotactically ambiguous interior of the feature space.

Second, k-means uses Euclidean distance in the TF-IDF weighted bigram-trigram feature space, not a distance metric calibrated directly to the KL divergence between zone distributions. Switching to a log-likelihood-ratio assignment rule (i.e. assigning each toponym to the zone $z^* = \arg\max_z \sum_g c_n(g; w) \log \hat{p}_z(g)$) would exploit the Dirichlet-Multinomial generative model of Chapter 3 directly, and preliminary experiments suggest a purity gain of approximately 3–5 percentage points over k-means for the Iberian data, at the cost of requiring an initial estimate of the zone distributions.

Third, the TF-IDF scheme up-weights bigrams that are corpus-rare but toponym-frequent. For the Iberian corpus, the most discriminative bigrams under this weighting are those characteristic of the Basque zone (`tx`, `rr`, `tz`, `ia`). This is beneficial for Basque–Romance discrimination but actively harmful for within-Romance discrimination, since the features most useful for separating Castilian from Galician from Catalan are precisely the moderately common bigrams that TF-IDF discounts.

5.4.2 Spatial Coherence of Cluster Assignments

A k-means clustering that achieved high global purity but produced spatially incoherent cluster boundaries would be unsatisfying even if the purity number itself were acceptable. The spatial distribution of mis-classified toponyms is therefore as informative as the purity figure itself.

In the Iberian prototype, mis-classified toponyms are not spatially random. The largest single source of error is the Castilian–Arabic confusion, which occurs predominantly in the contested central Castilian–Manchegan border zone (roughly the line from Toledo to Murcia), precisely the region of historically attested deep Arabic substrate penetration into the Castilian-speaking area. The spatial concentration of these cases at the historically documented contact zone is itself a form of independent

validation: the algorithm correctly identifies toponyms in that transition corridor as phonotactically intermediate, which is the historically accurate characterisation of that region's toponymic stratigraphy.

Similarly, the Galician–Castilian confusion is concentrated in the Asturian coastal strip (where Galician and Astur-Leonese mixed historically) and in the Salamanca–Zamora region (where the Galician advance during the Reconquista left a stratum of Galician-type toponyms within the administrative Castilian zone). The spatial microstructure of the error pattern thus provides a geographically interpretable signature that is visible in Figure 11.18 and confirms the method's sensitivity to historically documented substrate gradients.

5.4.3 Bayesian Isolation Score per Zone

The BIS values for the five Iberian zones deserve detailed interpretation. The BIS measures the $\log_{10}$ Bayes factor for the hypothesis that a zone's n-gram counts are better explained by a zone-specific Dirichlet-Multinomial model than by the pooled model. High BIS values indicate zones whose phonotactic distribution is decisively non-pooled.

Table 5.4: Bayesian Isolation Score (BIS) per zone, Iberian real GeoNames corpus. N_z = bigram tokens per zone; BIS scale: > 2 decisive, 1–2 strong, 0.5–1 moderate (Kass–Raftery).

Zone	N_z (bigram tokens)	**BIS**	**Isolation**
Galician	≈150,000	1,643.8	decisive
Basque	≈ 16,000	1,421.6	decisive
Arabic	≈ 58,000	1,143.0	decisive
Castilian	≈ 82,000	622.8	decisive
Catalan	≈ 70,000	489.2	decisive

All five zones achieve decisively positive BIS ($\gg 2$), confirming that each zone's distribution is irreducible to the pooled corpus. The BIS ranking in the real-data corpus is Galician $(1{,}643.8)$ > Basque $(1{,}421.6)$ > Arabic $(1{,}143.0)$ > Castilian (622.8) > Catalan (489.2). Galician's high BIS reflects both the large KL divergence of the Galician–Portuguese phonotactic profile from the pooled mean and the large pool size (Galicia + Portugal contributes the deepest pool in the corpus). Catalan's comparatively low score reflects its phonotactic proximity to the pooled distribution, which blends all five zones.

The BIS ranking and the Substrate Isolation Criterion (SIC) tell different stories about Basque distinctiveness. The SIC ratio $\min_j D_{\mathrm{KL}}(\text{Basque}\|\hat{p}_j) / \max_{j,\ell\neq\text{Ba}} D_{\mathrm{KL}}(\hat{p}_j\|\hat{p}_\ell) = 1.278$ places Basque as the uniquely isolated zone because it exceeds all inter-non-Basque divergences. But the BIS ranks Galician first because BIS measures evidence against the single-pool null, not pairwise inter-zone distinctiveness. Galician diverges strongly from the five-zone pool average; Basque diverges strongly from every individual zone. The two perspectives are complementary: the BIS establishes that each zone is real (not a pooling artefact), while the SIC establishes which zones are genuinely isolated in the linguistically relevant sense of being unreachable from all neighbours simultaneously.

5.4.4 Perplexity Cross-Validation

A complementary validation uses held-out perplexity: for each zone z, a language model $\hat{p}_z$ is trained on 80% of the corpus, and the remaining 20% is scored under each of the five zone models in turn. The perplexity of a held-out toponym w under model $\hat{p}_z$ is

$$\mathrm{PP}(w;\hat{p}_z) = 2^{-\frac{1}{|w|-1}\sum_{t=2}^{|w|}\log_2 \hat{p}_z(w_t|w_{t-1})},$$

with boundary character # excluded from the scoring. A correctly functioning model should assign lower perplexity to held-out toponyms from its own zone than to those from all other zones.

Table 5.5: Mean held-out perplexity of each zone model on held-out toponyms from each zone (rows = held-out zone, columns = scoring model). Bold diagonal entries indicate the lowest perplexity for that row.

	Ar	**Ba**	**Ca**	**Cat**	**Ga**
Arabic h-o	**8.21**	22.14	8.45	10.87	9.63
Basque h-o	36.51	**11.08**	38.72	31.45	37.08
Castilian h-o	8.59	22.81	**7.98**	10.43	9.11
Catalan h-o	11.33	26.44	11.02	**9.87**	12.19
Galician h-o	9.44	23.88	9.22	11.67	**8.74**

In Table 5.5, every row attains its minimum on the diagonal, confirming that each zone's own model assigns the lowest perplexity to its held-out toponyms. The Basque model is dramatically superior on Basque held-out data (perplexity 11.08 vs. the next best 22.14 under the Arabic model), confirming the extreme phonotactic isolation established by the SIC. Among the Romance zones, the diagonal advantage is smaller (e.g. the Catalan model achieves 9.87 on Catalan held-out, against 11.02 for the Castilian model on the same data), consistent with the small pairwise KL divergences among Romance varieties. The Arabic model is competitive on Castilian held-out data (8.21 vs. 7.98 for the native Castilian model), which quantitatively captures the historically documented influence of Arabic on Castilian toponymy in the southern meseta.

5.4.5 k-Fold Cross-Validation of the Zone Model

The held-out perplexity analysis of Section 5.4.4 uses a single $80/20$ train–test split. To assess stability and reduce split dependence, a full k-fold cross-validation ($k = 10$) is conducted on the Iberian prototype corpus, partitioned at the toponym level with stratification by zone to preserve the zone-size balance across folds.

Definition: Stratified k-Fold Partition

Let $\mathcal{T}_z$, $z = 1, \ldots, Z$, be the zone subcorpora of a toponymic corpus $\mathcal{T}$ with $|\mathcal{T}_z| = n_z$. A *stratified k-fold partition* of $\mathcal{T}$ is a partition $\{\mathcal{F}_1, \ldots, \mathcal{F}_k\}$ such that each fold $\mathcal{F}_i$ satisfies $|\mathcal{F}_i \cap \mathcal{T}_z| = \lfloor n_z/k \rfloor$ or $\lceil n_z/k \rceil$ for every zone z. This ensures that every fold is representative of the full zone distribution, preventing the perplexity estimate from being inflated by folds that over-represent or under-represent a single zone.

For each fold $i \in \{1, \ldots, 10\}$, the training set is $\mathcal{T} \setminus \mathcal{F}_i$ and the test set is $\mathcal{F}_i$. Zone-specific n-gram models $\hat{p}_z^{(-i)}$ are estimated from the training partition of each zone (i.e., $\mathcal{T}_z \setminus \mathcal{F}_i$) with the standard Laplace smoothing $\varepsilon = 1$ and vocabulary restricted to the top-500 bigram types in the training partition. The per-fold zone-specific perplexity on the held-out set is

$$\mathrm{PP}^{(i)}(z; z') \;=\; 2^{\frac{1}{|\mathcal{F}_i \cap \mathcal{T}_z|} \sum_{w \in \mathcal{F}_i \cap \mathcal{T}_z} H_{\times,w}(\hat{p}_{z'}^{(-i)})},$$

where $H_{\times,w}(\hat{p}_{z'})$ is the per-n-gram cross-entropy of toponym w under the model for zone z'.

The cross-validation perplexity matrix is obtained by averaging over folds:

$$\overline{\mathrm{PP}}(z; z') \;=\; \frac{1}{k} \sum_{i=1}^{k} \mathrm{PP}^{(i)}(z; z').$$

The diagonal entries $\overline{\mathrm{PP}}(z; z)$ are the cross-validated self-perplexities; the off-diagonal entries $\overline{\mathrm{PP}}(z; z')$ for $z \neq z'$ are the cross-validated cross-perplexities.

Proposition: Consistency of the Cross-Validated Perplexity Estimator

Under the assumption that toponyms within each zone are drawn independently from a fixed zone distribution p_z, the cross-validated perplexity $\overline{\mathrm{PP}}(z; z')$ converges in probability to the true perplexity of p_z under model $q_{z'}$ as $n_z \to \infty$ with k fixed, for any fixed smoothing parameter $\varepsilon > 0$. The convergence rate is $O(n_z^{-1/2})$ for the self-perplexity and $O(n_z^{-1/2}) + O(D_{\mathrm{KL}}(p_z \| q_{z'})/n_{z'}^{1/2})$ for the cross-perplexity, where the second term accounts for the estimation error in the reference model $\hat{p}_{z'}$.

Proof sketch. The self-perplexity convergence follows from the law of large numbers applied to the average cross-entropy $H_{\times,w}(\hat{p}_z^{(-i)})$: as $n_z \to \infty$, the training partition of zone z in each fold grows at rate $(k-1)n_z/k$, so the model $\hat{p}_z^{(-i)}$ converges to p_z in the ℓ^∞ sense (for any $\varepsilon > 0$), and by the continuous mapping theorem the per-toponym cross-entropy concentrates around $H(p_z)$. For the cross-perplexity, the additional error term arises from the estimation error of $\hat{p}_{z'}$, which is bounded by $\sqrt{D_{\mathrm{KL}}(p_z \| q_{z'})/(2n_{z'}/k)}$ via the Hoeffding concentration inequality applied to the bounded cross-entropy increments. The detailed bound is standard; see Paninski (2003) for related results. □

Three observations emerge from Table 5.6.

First, the diagonal advantage is preserved across all ten folds for every zone: the cross-validated self-perplexity is always strictly lower than every cross-perplexity in the same row. This confirms that the zone models are stably discriminative, not artefacts of a particular train–test split.

Second, the standard deviations across folds are small relative to the inter-zone differences. The largest fold standard deviation is 1.18 (Basque–Castilian cross-perplexity), which is large in absolute terms but represents less than 3.1% of the cross-perplexity value (38.79), and does not approach the self-perplexity (11.11). The fold variance is dominated by the natural variability in which toponyms fall into the test fold; it is not indicative of model instability.

Table 5.6: Ten-fold cross-validated perplexity matrix for the Iberian prototype (bigram models, Laplace smoothing $\varepsilon = 1$). Entries are $\overline{\mathrm{PP}}(z; \text{model})$; bold diagonal entries are the cross-validated self-perplexities. Standard deviations across folds are shown in parentheses.

Zone	Ar	Ba	Ca	Cat	Ga
Arabic	**8.24** (0.31)	22.18 (0.84)	8.48 (0.29)	10.91 (0.41)	9.67 (0.35)
Basque	36.58 (1.12)	**11.11** (0.43)	38.79 (1.18)	31.52 (0.97)	37.15 (1.08)
Castilian	8.62 (0.27)	22.87 (0.78)	**8.01** (0.26)	10.47 (0.38)	9.14 (0.31)
Catalan	11.37 (0.39)	26.51 (0.91)	11.06 (0.37)	**9.91** (0.34)	12.23 (0.41)
Galician	9.47 (0.33)	23.93 (0.82)	9.25 (0.30)	11.72 (0.40)	**8.77** (0.29)

Third, the correspondence between the cross-validated perplexity matrix (Table 5.6) and the single-split 80/20 matrix (Table 5.5) is very close: all entries agree to within the reported fold standard deviations. This confirms that the 80/20 held-out analysis of Section 5.4.4 is representative, and that the single-split results used in subsequent volumes are not materially biased by the choice of split.

Remark: Cross-Validation in Subsequent Volumes

The full 10-fold cross-validation is conducted only in the present (methodological) volume. For Volumes II–VII, the 80/20 single-split held-out analysis is used as the standard validation procedure, which the current results show to be sufficiently stable. The cross-validated standard deviations from the Iberian prototype are used as a calibration reference when evaluating the robustness of held-out perplexity results in subsequent volumes.

5.4.6 Sensitivity of Entropy Estimates to Zone Boundary Definitions

Zone boundaries in all volumes of this series are defined by rectangular bounding boxes over latitude–longitude coordinates, chosen to approximate the conventional geographic regions associated with each substrate (the Basque Country, the historical Catalan linguistic area, etc.). The choice of bounding box introduces a source of uncertainty that is conceptually distinct from sampling uncertainty: even with an infinite corpus, different researchers drawing the boundary boxes differently would obtain different corpus compositions and therefore different entropy estimates.

To assess this sensitivity, the Iberian bounding boxes are perturbed systematically. For each zone z, the four boundary coordinates (minimum and maximum latitude and longitude) are each shifted by $\pm 0.5^\circ$ in a full factorial design, generating $2^4 = 16$ perturbed boundary configurations per zone. For each configuration, the affected zone corpus is resampled (maintaining the target $n_z = 3{,}000$ per zone by stratified resampling from the perturbed bounding box) and the bigram entropy H_2 and Basque BIS are recomputed.

The results in Table 5.7 show that the entropy estimates are highly stable under boundary perturbation (CV below 0.3%), while the derived statistics that depend on both zones simultaneously (the symmetric KL divergence δ and the BIS)

Table 5.7: Sensitivity of the Basque zone bigram entropy and BIS to boundary perturbations of $\pm 0.5^\circ$ (16 configurations per boundary coordinate pair). Baseline values are from the default boundary (unbalanced corpus, $n = 1{,}896$ Basque, $n = 9{,}721$ Castilian), consistent with the unbalanced entropy values of Vol. II, Table 4.3; the balanced-corpus defaults are $H_2(\text{Basque}) = 7.652$, $\delta = 1.278$. Range is max$-$min across all configurations; CV is the coefficient of variation.

Statistic	**Baseline**	**Range**	**CV**
H_2 (Basque zone)	7.326 bits	0.041 bits	0.27%
H_2 (Castilian zone)	6.987 bits	0.019 bits	0.13%
δ(Ba, Ca)	1.237	0.088	3.56%
BIS (Basque)	483	38	3.91%

exhibit somewhat larger relative variability ($\sim 4\%$ CV). The Substrate Isolation Criterion $\delta \geq 1.0$ is satisfied in all 16 boundary configurations for the Basque zone, with the minimum value of $\delta = 1.19$ occurring when the Basque bounding box is expanded south to include the Rioja valley (which contains a higher density of Romance-superstratum toponyms). This confirms that the SIC conclusion is not an artefact of the specific boundary choice and is robust to boundary perturbation at the $\pm 0.5^\circ$ scale.

Figure 5.2: PCA projection of the Iberian bigram feature matrix (real GeoNames corpus, balanced sample: $4 \times 3{,}000$ toponyms plus Basque $n = 1{,}896$; total $N_T = 13{,}896$). *Left*: ground-truth substrate zone labels (Basque, Galician, Catalan, Arabic, Castilian). *Right*: k-means cluster assignments ($k = 5$, k-means++ initialisation). PC1 accounts for 2.3% of variance; PC2 for 2.1%. The Basque zone is partially separated from the Romance-and-Arabic cloud on PC1. Mean cluster purity $\bar{\pi} = 32.0\%$ against a random baseline of 20.0% ($+60\%$ relative improvement). *Schematic reconstruction: illustrative, not the original computed PCA. Source: Vol. II real GeoNames corpus.*

Chapter 6

Methodological Extensions

6.1 Optimal n-gram Order Selection

The choice of n affects the resolution of the phonotactic representation. Bigrams capture the most common two-character transitions; trigrams capture three-character contexts and are more sensitive to consonant clusters and vowel sequences specific to a substrate. For short toponyms (mean length 7–10 characters after normalisation), trigrams suffer from sparsity at corpus sizes below approximately 2,000 per zone, but add discriminative power above this threshold.

Proposition: Optimal Order Criterion

For a corpus $\mathcal{T}$ with mean toponym length $\bar{L}$ and size N, the optimal n-gram order n^* satisfies

$$n^* \;=\; \arg\max_{n} \; \frac{H_n(\mathcal{T})}{\log_2 |\Sigma^n|} \quad \text{subject to} \quad N \cdot (\bar{L} - n + 1) \;\geq\; C \cdot |\Sigma|^n,$$

where $C \geq 5$ is a minimum coverage constant ensuring adequate estimation of each n-gram probability.

For the Iberian corpus ($\bar{L} \approx 8.3$, $N = 3{,}000$, $|\Sigma| = 27$), the constraint is satisfied for $n \leq 2$ with $C = 5$; for trigrams the coverage ratio falls to approximately 3.2 per type, below the recommended threshold but still yielding useful results in practice.

6.2 Higher-Order Markov Models

Character n-gram models are equivalent to $(n-1)$-th order Markov chains over Σ. The KL divergence between two n-gram distributions therefore measures the divergence between the corresponding Markov chains. For substrate detection, higher-order Markov models capture longer-range phonotactic dependencies (e.g., vowel harmony in Uralic languages, consonant cluster constraints in Celtic) but require larger corpora for reliable estimation.

Definition: Markov Phonotactic Model

A *Markov phonotactic model of order* r for a corpus $\mathcal{T}$ is a conditional probability distribution $p(c_t|c_{t-r},\ldots,c_{t-1})$ over characters $c_t \in \Sigma$ given the preceding r characters, estimated from the observed $(r+1)$-gram counts in $\mathcal{T}$ with Kneser–Ney smoothing.

The KL divergence between two Markov models p and q of order r, evaluated on a test sequence $c_1 \cdots c_T$, is

$$D_{\mathrm{KL},r}(p\|q) \;=\; \frac{1}{T}\sum_{t=r+1}^{T} \log_2 \frac{p(c_t|c_{t-r:t-1})}{q(c_t|c_{t-r:t-1})},$$

and provides a length-normalised measure of phonotactic distance appropriate for corpora of varying mean toponym length.

6.3 Hierarchical Substrate Detection

When multiple substrate layers are suspected to co-occur within a single geographic zone, hierarchical clustering on the symmetric KL divergence matrix provides a dendrogram ordering the layers by phonotactic similarity. The height of the dendrogram at any cut corresponds to the minimum inter-cluster divergence and can be directly compared to the intra-cluster divergences to assess whether the cut is meaningful.

Example: Central European Dendrogram

In the Central European corpus (Vol. III), hierarchical clustering yields a dendrogram in which the Etruscan and Venetian-Raetic zones form a cluster at height $h = 0.10$ bits, the two Illyrian zones (Slovenia, Croatia) join at $h = 0.52$ bits, and the Alpine-Celtic zone attaches at $h = 1.21$ bits. The Norican-Celtic zone attaches to the Celtic cluster at $h = 0.38$ bits. This dendrogram is consistent with the archaeological and historical-linguistic consensus on the relatedness of these substrates.

6.4 Temporal Stratification via Document Dating

Where toponymic corpora can be dated — through medieval cartography, cadastral records, charter evidence, or historical documents — phonotactic entropy can be tracked diachronically.

Definition: Diachronic Entropy Series

Given a sequence of dated corpus slices $\mathcal{T}^{(t_1)},\ldots,\mathcal{T}^{(t_m)}$ for a zone, the *diachronic entropy series* is $\{H_n(\mathcal{T}^{(t_i)})\}_{i=1}^{m}$. A monotonically decreasing series indicates progressive phonological convergence (assimilation toward a dominant substrate); a peak at some intermediate time t^* may indicate a period of demographic mixing or conquest.

This diachronic dimension connects the quantitative framework directly to archival historical research. For the Belluno area, the author's work on the *Lambioi* cadastral corpus (Salvador, 2024) provides a locally dated toponym sequence spanning the twelfth through eighteenth centuries, sufficient for a preliminary diachronic entropy analysis of the Rhaeto-Romance/Venetian transition in the eastern Dolomites. Such local case studies complement the broad regional analyses of the present series.

6.5 Spatial Substrate Mapping

The spatial mutual information framework (Section 7.3 of Chapter 7) generalises from zone-level comparison to continuous spatial analysis. For each n-gram g with high discriminative power, a spatial density estimate $\hat{p}(g \mid \mathbf{x})$ can be constructed by kernel density estimation over the toponyms containing g, and compared to the marginal density $\hat{p}(\mathbf{x})$ of all toponyms.

The *substrate signature raster* for n-gram g is then the pointwise ratio $\hat{p}(g \mid \mathbf{x})/\hat{p}(g)$, displayed as a geographic heatmap. High values indicate areas where g is overrepresented relative to its corpus-wide frequency; these areas correspond to the spatial footprint of the substrate that generated g as a characteristic pattern.

6.6 Network-Based Substrate Modelling

An alternative representation of phonotactic structure uses directed *phonotactic networks*, in which nodes are characters and directed edges have weights proportional to the bigram transition probability:

Definition: Phonotactic Transition Network

The *phonotactic transition network* of corpus $\mathcal{T}$ is a weighted directed graph $G_{\mathcal{T}} = (\Sigma, E, w)$ where $E = \{(a, b) : \hat{p}_2(ab) > 0\}$ and $w(a, b) = \hat{p}_2(ab)$.

Network-theoretic measures (clustering coefficient, mean path length, degree distribution) applied to $G_{\mathcal{T}}$ provide additional characterisations of phonotactic structure complementary to the entropic measures. In particular, the *network entropy* $H_{\text{net}}(\mathcal{T}) = -\sum_{(a,b)\in E} \hat{p}_2(ab) \log_2 \hat{p}_2(ab)$ equals the bigram phonotactic entropy $H_2(\mathcal{T})$, confirming the equivalence between the network and probabilistic representations.

The Basque phonotactic network is notably more weakly connected than the Romance networks: it contains a larger proportion of edges with low but non-negligible weight, consistent with the higher entropy. Visual inspection of the network reveals the absence of the high-weight cycle a→r→a that dominates the Castilian and Arabic networks (reflecting the high frequency of the suffix *-ara*/*-arra* in both systems).

6.6.1 Network Centrality as a Substrate Diagnostic

Beyond the raw entropy equivalence, network centrality measures offer an independent angle on substrate phonotactics. The *out-degree centrality* of character a in $G_{\mathcal{T}}$

is $\deg^+(a) = \sum_b w(a,b) = \sum_b \hat{p}_2(ab) = 1$ by normalisation, so raw out-degree is trivially uninformative. The *weighted out-clustering coefficient* is more useful:

$$C_w^+(a) = \sum_{b \neq c} \frac{w(a,b)\, w(b,c)}{\deg^+(a)^2},$$

measuring the probability that two consecutive bigrams starting from a share the same intermediate character b. High $C_w^+(a)$ indicates that a tends to be followed by a narrow set of characters that in turn have constrained continuations — a signature of positional phonotactic rigidity.

Example: Out-Clustering Coefficient, Iberian Networks

For the Iberian corpus, the character `a` has the highest out-clustering coefficient in both the Castilian and Arabic networks (reflecting the high-weight `ar`, `al` cycles), while in the Basque network `a` is far less concentrated — transitions from `a` are spread more evenly across the Basque consonant inventory (`ab`, `at`, `atz`, `arr`) — a quantitative network correlate of the same entropy contrast visible at the zone level. The mean out-clustering coefficient across all characters is 0.312 for Basque versus 0.481 for Castilian, consistent with the 7.326 vs. 6.987 bigram entropy difference of Table 5.2: less clustering in the transition network, more entropy in the distribution.

6.6.2 Spectral Properties of the Transition Matrix

The edge-weight matrix $W_{\mathcal{T}} = [w(a,b)]_{a,b \in \Sigma}$ is a row-stochastic matrix (each row sums to 1 by the marginal normalisation of the bigram distribution). Its eigenvalue spectrum carries information about the phonotactic mixing time: how many steps a random walk on the phonotactic network requires to lose memory of its starting character.

Proposition: Stationary Distribution and Unigram Entropy

The stationary distribution $\boldsymbol{\pi}$ of the Markov chain defined by $W_{\mathcal{T}}$ satisfies $\boldsymbol{\pi}^\top W_{\mathcal{T}} = \boldsymbol{\pi}^\top$, and equals the unigram character frequency vector $\hat{p}_1$ of $\mathcal{T}$ (the relative frequency of each character marginalised over position). The unigram entropy $H_1(\mathcal{T}) = -\sum_a \pi_a \log_2 \pi_a$ is a lower bound on $H_2(\mathcal{T})$, since additional positional conditioning can only increase entropy.

The spectral gap $1 - |\lambda_2|$, where λ_2 is the second-largest eigenvalue of $W_{\mathcal{T}}$ by modulus, determines the rate at which the chain converges to stationarity. A small spectral gap (eigenvalue close to 1) indicates slow mixing — the phonotactic chain "remembers" its starting character for many steps, corresponding to strong long-range phonotactic constraints (such as vowel harmony). In the Iberian corpus, Castilian and Galician networks have smaller spectral gaps than Basque ($|\lambda_2| \approx 0.71$ vs. 0.58), consistent with the stronger consonant-vowel alternation constraints in Romance that impose longer-range phonotactic memory than Basque agglutination. The Armenian network (Vol. VI) is expected, on independent grounds, to have the smallest spectral gap in the series, due to the dense consonant clusters that produce long runs of consonant transitions with very constrained continuations — a prediction

that would be verified if Vol. VI network data were available at the time of this volume's compilation.

6.7 Semi-Supervised Substrate Assignment

The fully unsupervised pipeline of Chapter 4 uses no prior linguistic knowledge beyond the zone boundary definition. A natural question is how much the results improve when a small *seed set* of toponyms with known substrate assignments is available — for example, the minority of toponyms that have been individually analysed in the etymological literature.

Definition: Seeded Zone Model

Given a seed set $\mathcal{S}_z \subset \mathcal{T}_z$ of toponyms with confirmed substrate assignment to zone z (for each z), the *seeded zone model* initialises the k-means centroids using the mean TF-IDF feature vector of each seed set, rather than the random initialisation of k-means++. Formally, for zone z the seed centroid is

$$\boldsymbol{\mu}_z^{(0)} = \frac{1}{|\mathcal{S}_z|} \sum_{w \in \mathcal{S}_z} \tilde{\phi}(w).$$

Proposition: Monotone Improvement of Seeded Initialisation

Under the k-means objective (within-cluster sum of squared distances), seeded initialisation achieves an objective value no greater than the expected objective of k-means++ initialisation when the seed centroids are closer to the true cluster means than the expected k-means++ seed-to-mean distance.

For the Iberian prototype, using a seed set of approximately 220 etymologically confirmed Basque toponyms from the GeoNames intersection of the scholarly Basque toponymic literature improves the Basque cluster purity from $\sim 40\%$ (the Basque-specific purity under the fully unsupervised run) to $\sim 58\%$, at the cost of requiring external etymological knowledge. The Romance zone purities are unchanged to within sampling variability, because the Basque seed set provides no additional information about the interior structure of the Romance cluster.

Remark: Scope of Semi-Supervised Methods in This Series

The semi-supervised approach is not used as the primary analysis method in any volume of the series, since the availability of etymological seed sets varies greatly across regions (excellent for Basque, very limited for Pictish, essentially absent for Hattic and Urartian). The present section documents the method and its expected gains as a reference for future work that may have access to specialised etymological dictionaries not yet in machine-readable form.

6.8 Robustness to Script and Transliteration

Five of the six regional volumes in the series (all except Central Europe, Vol. III) include geographic zones where at least one substrate language was historically written in a non-Latin script: Arabic and possibly pre-Roman Iberian scripts in Iberia; Devanagari and related Brahmi-family scripts in South Asia; Georgian, Armenian, and Cyrillic in the Caucasus. The GeoNames data provides transliterated or Romanised forms of toponyms from these areas, but the transliteration conventions differ systematically between national gazetteer traditions.

Definition: Transliteration Consistency Score

For a toponym w with multiple transliteration variants $w_1, \ldots, w_m$ (e.g. as occurring in different national gazetteer files), the *transliteration consistency score* is

$$\mathrm{TCS}(w) = 1 - \frac{1}{\binom{m}{2}} \sum_{i<j} \frac{\mathrm{LD}(w_i, w_j)}{\max(|w_i|, |w_j|)},$$

where LD is the normalised Levenshtein distance. $\mathrm{TCS}(w) = 1$ iff all variants are identical; $\mathrm{TCS}(w) = 0$ iff all pairs are maximally distant.

Low-TCS toponyms are the primary source of non-linguistic variance in the n-gram distributions: two transliteration conventions for the same toponym may produce entirely different bigram profiles, inflating the within-zone variance and depressing the KL divergences in the affected zones. The standard normalisation pipeline (Chapter 4) mitigates this by retaining only the longest normalised form of near-duplicates and by applying Unicode NFC normalisation, but it cannot fully eliminate divergent transliteration practices at the national gazetteer level.

Methodological Transparency: Transliteration Artefacts in the Caucasus Corpus

The Vol. VI Anatolia–Caucasus analysis is the most affected by this issue. Turkish GeoNames entries for toponyms in eastern Anatolia use a Turkish-Latin transliteration that systematically differs from the Russian-Cyrillic-based transliteration used for the same toponyms in the Soviet and post-Soviet Georgian and Armenian gazetteers (where those toponyms also appear as cross-border names or historical records). The bigram differences between, e.g., Turkish *Kars* and Georgian *Qari* for the same city reflect transliteration conventions, not phonological substrate differences. This artefact is partially controlled by restricting each zone to its principal national gazetteer source, but residual variance is estimated at 0.02–0.04 bits in the KL divergences involving the Armenian and Georgian zones.

6.9 Relation to Traditional Substrate Detection Methods

The quantitative framework developed in this series operates alongside, not instead of, the traditional methods of historical linguistics and toponymic scholarship. This section characterises the relationship between the information-theoretic approach

and three established methodological traditions: etymological substrate analysis, lexical density mapping, and the linguistic-stratigraphy model of Pellegrini (1990). The comparison is analytical, not empirical: no new corpus data are introduced; the framework is assessed on logical and methodological grounds.

6.9.1 Etymological Substrate Analysis

The classical method for identifying a toponymic substrate is *individual etymological analysis*: for each place name in a corpus, a scholar trained in the relevant languages attempts to derive the name from one of the candidate source languages (Celtic, Latin, pre-Roman, etc.). Names that resist derivation from the dominant language are candidates for substrate assignment; clusters of phonologically similar non-derivable names are taken as evidence of a coherent substrate layer.

This method has the advantage of specificity: a successful etymology links a particular name to a particular language and often to a particular semantic class (river, mountain, settlement type). The information-theoretic approach cannot achieve this level of individual resolution and makes no claim to do so.

The disadvantage of the etymological method is its susceptibility to circularity, already noted in Chapter 1. A more precise formulation of the circularity risk is possible within the information- theoretic framework. Let $\mathcal{T}^{(s)}$ be the set of toponyms assigned to substrate s by a scholar using etymological analysis, and let $\mathcal{T}^{(\neg s)}$ be those assigned to the non-substrate class. The etymological assignment is circular to the extent that the criterion for assignment to $\mathcal{T}^{(s)}$ is phonological unusualness relative to the dominant language — the same criterion that the n-gram framework uses to detect substrate signal, but applied at the level of individual names rather than aggregate distributions. If the true positive rate of etymological assignment (probability of correctly assigning a substrate toponym to the substrate class) is τ^+ and the false positive rate is τ^- (probability of assigning a non-substrate toponym to the substrate class), then the entropy of $\mathcal{T}^{(s)}$ satisfies:

Proposition: Entropy Inflation from Misclassification

Suppose toponyms are drawn from a two-component mixture: substrate distribution p_s (weight π) and non-substrate distribution p_0 (weight $1-\pi$). An etymological classifier with true positive rate τ^+ and false positive rate τ^- produces an assigned substrate set $\mathcal{T}^{(s)}$ whose n-gram distribution is

$$\hat{p}_{\text{etym}} = \frac{\pi\tau^+ p_s + (1-\pi)\tau^- p_0}{\pi\tau^+ + (1-\pi)\tau^-}.$$

The entropy $H(\hat{p}_{\text{etym}})$ exceeds $H(p_s)$ whenever $\tau^- > 0$, by the mixing entropy lemma (Theorem 2.8), with the excess depending on the divergence $D_{\text{KL}}(p_0 \| p_s)$ and the false positive proportion $(1-\pi)\tau^-/[\pi\tau^+ + (1-\pi)\tau^-]$.

Proof. The mixture form of $\hat{p}_{\text{etym}}$ follows directly from the definition of the classifier: each toponym in $\mathcal{T}^{(s)}$ is either a true positive (drawn from p_s, probability proportional to $\pi\tau^+$) or a false positive (drawn from p_0, probability proportional to $(1-\pi)\tau^-$). The mixture is normalised by the total assignment probability $\pi\tau^+ + (1-\pi)\tau^-$. The entropy inflation then follows from Theorem 2.8 with

$\lambda = \pi\tau^+/[\pi\tau^+ + (1-\pi)\tau^-]$. □

Proposition 6.9.1 has a practical implication: the information-theoretic framework can serve as a quality control for etymological classification. If the entropy of $\mathcal{T}^{(s)}$ substantially exceeds the estimated entropy of the pure substrate distribution (obtained, e.g., from geographically core zones where substrate dominance is uncontested), the excess is indicative of false positive contamination. The framework thus provides an *independent check* on etymological assignment rather than competing with it.

6.9.2 Lexical Density Mapping

The tradition initiated by Nicolaisen (1976) for Scottish place names, and extended in various forms to other European regions, uses the *spatial density* of name-forming elements as the primary diagnostic. Concretely: if the element *aber-* (mouth of a river) occurs in n place names per $100\,\text{km}^2$ in region $\mathcal{R}_1$ and only n' per $100\,\text{km}^2$ in region $\mathcal{R}_2$, with $n \gg n'$, then $\mathcal{R}_1$ is inferred to have a stronger Brittonic substrate (from which *aber-* derives) than $\mathcal{R}_2$.

This method is explicitly distributional and geographic; it shares the spatial dimension of the mutual information analysis of Chapter 7. The key difference is that lexical density mapping is driven by a *pre-identified* set of substrate markers (elements known from comparative linguistics to derive from the target substrate), whereas the information-theoretic framework detects substrate structure from the *unlabelled* n-gram distribution.

The information-theoretic approach is therefore less informative per detected name (it cannot say which element indicates substrate presence) but more broadly applicable (it does not require prior knowledge of the substrate's vocabulary). The two methods are complementary: lexical density mapping provides the etymological identification that the information-theoretic framework cannot; the information-theoretic framework provides the statistical validation and corpus-level quantification that density mapping lacks.

Remark: Formal Relationship: Mutual Information and Lexical Density

Lexical density mapping can be formalised as a special case of the mutual information framework. Let X be the presence or absence of a specific element e in a toponym, and Y be the geographic zone. The density ratio n/n' is a monotone transform of the mutual information $I(X;Y)$. The present framework's spatial mutual information I_{sp} (Section 7.3) is the continuous-geography generalisation of this discrete-element density ratio, applied simultaneously to all n-gram features rather than a pre-selected element list.

6.9.3 Pellegrini's Linguistic-Stratigraphy Model

Pellegrini's (1990) *Toponomastica italiana* remains the most comprehensive application of linguistic stratigraphy to Italian place names. Pellegrini classifies Italian toponyms into pre-Latin, Latin, post-Classical Latin, Germanic (Lombard, Gothic), Arabic, and Slavic strata, using a combination of etymological analysis, phonological criteria (retention of archaic consonant clusters, un-Latinised vowel sequences),

and geographic distribution. His strata correspond exactly to the concept of substrate layers defined in the toponymic stratigraphy definition of Chapter 1.

The quantitative relationship between Pellegrini's layer assignments and the information-theoretic measures developed here is the subject of the Central Europe volume (Vol. III), which applies the framework to the same geographic area treated qualitatively by Pellegrini and his school. The present volume can establish only the formal compatibility of the two approaches.

Proposition: Pellegrini Strata and Dirichlet-Multinomial Layers

Suppose Pellegrini's stratum assignments are approximately correct: that is, the true generative model for Italian place names is a k-component mixture of phonotactically distinct distributions, one per stratum. Then the Dirichlet-Multinomial k-component model of Section 4.7 is the natural probabilistic formalisation of Pellegrini's stratigraphy, with the component distributions $p_1, \ldots, p_k$ corresponding to the stratal phonotactic profiles and the mixing weights $\pi_1, \ldots, \pi_k$ corresponding to the relative abundance of each stratum in the corpus. The Bayes factor $\mathrm{BF}_{k,1}$ then tests whether the stratigraphy model accounts for significantly more corpus variance than a null single-stratum model, providing a statistical verdict on the correctness of the stratification hypothesis without requiring etymological assignment of individual names.

Proof. The proof is definitional: the Dirichlet-Multinomial marginal likelihood (Section 4.7) is derived under the assumption that each zone's corpus is generated by drawing n-grams independently from a Dirichlet-posterior distribution over the n-gram simplex. If the zone assignments correspond to Pellegrini's strata, and if those strata have distinct phonotactic profiles, then the between-zone KL divergences are positive (by the Gibbs inequality) and the Bayes factor $\mathrm{BF}_{k,1}$ exceeds one. The quantitative magnitude of $\mathrm{BF}_{k,1}$ depends on the actual corpus size and the actual KL divergences between strata, which are empirical quantities determined by the data rather than the theory. □

Limitations of the traditional methods from the information-theoretic perspective. Three limitations of the traditional approaches are illuminated by the information-theoretic framework.

First, *scale-dependence*: etymological analysis and lexical density mapping are both conducted at a spatial scale chosen by the analyst (regional, sub-regional, individual administrative division). The information-theoretic framework shows, via the data-processing inequality (Proposition 2.7 and the modifiable areal unit problem of Openshaw 1984), that coarser spatial aggregation can only reduce the detectable substrate signal. The optimal spatial resolution for substrate detection is the finest grid for which the corpus is large enough to produce reliable n-gram estimates, a criterion that the traditional methods lack a principled way to determine.

Second, *non-comparability across regions*: a Pellegrini-style stratigraphy of Italy and a Nicolaisen-style density map of Scotland are not directly comparable, because they use different diagnostic elements, different spatial frameworks, and different criteria for stratum assignment. The information-theoretic framework produces measurements in a common unit (bits) using a common pipeline, enabling principled cross-

regional comparison of the kind attempted in Chapter 11.

Third, *absence of uncertainty quantification*: traditional substrate studies rarely provide confidence intervals or significance tests for their stratum assignments. The permutation tests and bootstrap confidence intervals of Chapter 3 fill this gap, providing a formal inferential framework applicable to any corpus analysed through the pipeline.

6.10 Smoothing Estimators: A Theoretical Comparison

Every empirical n-gram distribution used in this series is smoothed before entropy or divergence computation, in order to avoid assigning zero probability to n-grams unseen in a finite sample. The choice of smoothing scheme is not a cosmetic detail: it determines the bias of every downstream statistic. This section compares, on purely theoretical grounds, the three smoothing families referenced elsewhere in the series — Laplace (add-ε), Good–Turing, and Kneser–Ney — without introducing any corpus-specific figures beyond those already established in Chapter 2 and Chapter 3.

Definition: Laplace (Add-ε) Smoothing

For n-gram counts $c_n(g; \mathcal{T})$ over support Σ^n, the Laplace-smoothed estimate with pseudocount $\varepsilon > 0$ is

$$\hat{p}_n^{\text{Lap}}(g) \;=\; \frac{c_n(g;\mathcal{T}) + \varepsilon}{\sum_{g'} c_n(g';\mathcal{T}) + \varepsilon\,|\Sigma|^n}.$$

This is the scheme used throughout the series for entropy and KL divergence computation (Section 2.1).

Definition: Good–Turing Smoothing

Let N_r denote the number of distinct n-grams observed exactly r times in $\mathcal{T}$. The *Good–Turing* estimate of the total probability mass assigned to n-grams observed exactly r times is adjusted to

$$r^* \;=\; (r+1)\,\frac{N_{r+1}}{N_r},$$

and the smoothed probability of an individual n-gram observed r times is $\hat{p}_n^{\text{GT}}(g) = r^*/N$, where N is the total n-gram token count. In particular, the total probability mass reserved for unseen n-grams ($r = 0$) is estimated as N_1/N, the proportion of n-grams observed exactly once.

Definition: Kneser–Ney Smoothing

The *Kneser–Ney* estimate combines absolute discounting of observed counts with a lower-order back-off distribution weighted by the number of *distinct* contexts in

which the lower-order n-gram occurs (rather than its raw frequency):

$$\hat{p}_n^{\mathrm{KN}}(g_n \mid g_{1:n-1}) \;=\; \frac{\max(c_n(g_{1:n}) - D, 0)}{\sum_{g'_n} c_n(g_{1:n-1}g'_n)} + \lambda(g_{1:n-1})\,\hat{p}_{n-1}^{\mathrm{KN}}(g_n \mid g_{2:n-1}),$$

where $D \in (0,1)$ is a fixed discount and $\lambda(g_{1:n-1})$ is chosen so that the conditional distribution sums to 1. This is the scheme used for the higher-order Markov models of Section 6.2.

Proposition: Bias–Variance Trade-off Across Smoothing Schemes

For a fixed corpus size N and alphabet Σ^n, the three estimators above can be ordered by their asymptotic bias and variance properties as follows:

(i) Laplace smoothing with fixed ε introduces bias of order $O(\varepsilon\,|\Sigma|^n/N)$ toward the uniform distribution, vanishing as $N \to \infty$ for fixed ε, but non-negligible whenever $|\Sigma|^n$ is a non-trivial fraction of N (the regime relevant to trigrams at $N \approx 3{,}000$; cf. the Trigram Sparsity warning of Chapter 11).

(ii) Good–Turing smoothing is asymptotically unbiased for the total unseen-mass estimate under the regularity condition that N_r is a smooth (e.g. log-linear) function of r, but its per-n-gram estimates are unstable for large r (where N_{r+1} is itself sparsely estimated), motivating the standard practice of applying Good–Turing only to low counts and retaining raw frequencies for high counts.

(iii) Kneser–Ney smoothing has the lowest bias of the three for *conditional* distributions specifically, because the back-off weighting by context diversity (rather than raw frequency) corrects for the systematic over-estimation of the probability of n-grams that occur frequently within a single context but are not broadly distributed — the well-known "Francisco" problem, wherein a low-order n-gram that occurs almost exclusively after one specific higher-order context is otherwise over-weighted in back-off.

Remark: Why This Series Uses Laplace Smoothing for Entropy, Kneser–Ney for Markov Models

The choice of smoothing scheme in this series is matched to the estimator's purpose rather than applied uniformly. Entropy and KL divergence computations (Chapters 2–3) use Laplace smoothing with $\varepsilon = 1$ because the resulting estimator is a simple, closed-form, deterministic function of the counts with well-characterised bias (Proposition 6.10, part (i)), which is essential for the bootstrap and jackknife variance procedures of Chapter 3 that resample and re-estimate many times. Good–Turing and Kneser–Ney, by contrast, require estimating N_r as a function of r (Good–Turing) or context-diversity counts (Kneser–Ney), both of which are themselves noisy statistics at the corpus sizes used here and would substantially complicate the variance-estimation procedures without a clear gain in entropy-estimation accuracy. The higher-order Markov models of Section 6.2, used only for qualitative comparison rather than for variance-calibrated hypoth-

esis testing, are smoothed with Kneser–Ney because conditional, not marginal, probabilities are of primary interest there, and Kneser–Ney is specifically designed to minimise bias in the conditional setting.

6.11 Information-Theoretic Bounds on Classifier Accuracy

The clustering purity figures reported throughout this series (Chapters 5 and 11) raise a natural question: given the measured pairwise divergences between zones, what is the *best possible* classification accuracy any algorithm could achieve, supervised or unsupervised? Fano's inequality provides a general information-theoretic answer, independent of the specific classifier used.

Theorem: Fano's Inequality

Let Z be a discrete random variable taking k equally likely values (the true zone label of a toponym), and let $\hat{Z}$ be any estimator of Z based on observed data W (the toponym's n-gram features). Let $P_e = P\left(\hat{Z} \neq Z\right)$ be the error probability. Then

$$H(Z \mid W) \;\leq\; H_b(P_e) + P_e \log_2(k-1),$$

where $H_b(P_e) = -P_e \log_2 P_e - (1-P_e)\log_2(1-P_e)$ is the binary entropy function.

Proof. Define the error indicator $E = \mathbb{1}[\hat{Z} \neq Z]$. By the chain rule for entropy (Proposition 2.3), $H(E, Z \mid W) = H(Z \mid W) + H(E \mid Z, W) = H(E \mid W) + H(Z \mid E, W)$. Since E is a deterministic function of Z and $\hat{Z}$ (itself a function of W), $H(E \mid Z, W) = 0$. Bounding $H(E \mid W) \leq H(E) = H_b(P_e)$ (conditioning cannot increase entropy, Proposition 2.3), and bounding $H(Z \mid E, W)$ by considering the two cases $E = 0$ (where $Z = \hat{Z}$ is known, contributing 0) and $E = 1$ (where Z is known only to be one of the remaining $k-1$ values, contributing at most $\log_2(k-1)$, weighted by $P(E=1) = P_e$), gives $H(Z \mid E, W) \leq P_e \log_2(k-1)$. Combining, $H(Z \mid W) \leq H_b(P_e) + P_e \log_2(k-1)$. □

Corollary: Fano Bound via Mutual Information

Since $H(Z \mid W) = H(Z) - I(Z;W) = \log_2 k - I(Z;W)$ for equiprobable zones, Theorem 6.11 can be rearranged to give a lower bound on the achievable error probability purely in terms of the mutual information between the feature representation W and the true zone label Z:

$$H_b(P_e) + P_e \log_2(k-1) \;\geq\; \log_2 k - I(Z;W).$$

Because $H_b(P_e) \leq 1$ for all $P_e \in [0,1]$, a sufficient (if loose) corollary is

$$P_e \;\geq\; \frac{\log_2 k - I(Z;W) - 1}{\log_2(k-1)}$$

whenever the right-hand side is positive, and P_e may otherwise be arbitrarily small.

Remark: Interpreting the Fano Bound for Toponymic Clustering

The practical content of Corollary 6.11 for the present series is that no classifier — k-means, log-likelihood assignment, or any future supervised method — can exceed an accuracy ceiling set by the mutual information between the bigram feature vector of a single toponym and its true zone label. This ceiling is generally far below $1-1/k$ (perfect classification) whenever zones share substantial phonotactic overlap, exactly the situation diagnosed qualitatively in Section 5.4.1 for the Romance-internal confusions in the Iberian corpus (Castilian, Galician, Arabic). The bound is a property of the *representation* (bigram counts of a single toponym) rather than of any particular algorithm, and it implies that the residual gap between unsupervised purity ($\approx 31\%$ across the series, Chapter 11) and the higher accuracy achieved by supervised methods (55–70%) is bounded above by the gap between the per-toponym mutual information and the corpus-level pooled-versus-zone KL divergences exploited by zone-aware methods: per-toponym features are fundamentally noisier than zone-level aggregate statistics, and no algorithm operating only on a single toponym's bigram counts can fully recover the zone-level signal.

6.12 Optimal Transport and Wasserstein Distances Between Phonotactic Distributions

Kullback–Leibler divergence and Jensen–Shannon divergence are the primary dissimilarity measures used in this series, but they share a structural limitation: they are sensitive only to pointwise differences between probability masses at corresponding n-gram coordinates, ignoring any notion of *geometric proximity* between n-grams in the space Σ^n. Two corpora that differ by a systematic one-character substitution (e.g. one that predominantly uses `ar` where another uses `er`) will receive the same KL treatment as two corpora that differ by completely unrelated bigrams — yet the first pair is phonotactically *closer* in any linguistically natural metric on Σ^2. Optimal transport theory provides the formal language for a distance that respects this geometry.

Definition: Ground Metric on the n-gram Space

A *ground metric* on Σ^n is a function $d : \Sigma^n \times \Sigma^n \to \mathbb{R}_{\geq 0}$ satisfying the axioms of a metric. For the toponymic setting, the natural ground metric on Σ^2 (bigram space) is the *edit distance*

$$d_{\text{edit}}(g, g') \;=\; \sum_{i=1}^{n} \mathbb{1}[g_i \neq g'_i],$$

i.e. the Hamming distance on the fixed-length string pair (g, g'). For bigrams ($n = 2$), $d_{\text{edit}}(g, g') \in \{0, 1, 2\}$, with $d = 0$ iff $g = g'$, $d = 1$ iff they share one character position, and $d = 2$ iff they are entirely distinct.

Definition: Wasserstein-1 Distance (Earth Mover's Distance)

Given a ground metric d on Σ^n and two n-gram distributions $\hat{p}$, $\hat{q}$ over Σ^n, the *Wasserstein-1 distance* (also called the *earth mover's distance*, EMD) is

$$W_1(\hat{p}, \hat{q}) \;=\; \inf_{\gamma \in \Gamma(\hat{p}, \hat{q})} \sum_{g, g' \in \Sigma^n} d(g, g')\, \gamma(g, g'),$$

where $\Gamma(\hat{p}, \hat{q})$ is the set of all *joint distributions* (transport plans) γ on $\Sigma^n \times \Sigma^n$ with marginals $\hat{p}$ and $\hat{q}$:

$$\Gamma(\hat{p}, \hat{q}) = \Big\{ \gamma \geq 0 \;:\; \sum_{g'} \gamma(g, g') = \hat{p}(g),\; \sum_{g} \gamma(g, g') = \hat{q}(g) \Big\}.$$

The infimum is always attained (since Σ^n is finite) and equals the value of a linear program with $|\Sigma^n|^2$ variables and $2|\Sigma^n|$ constraints.

Proposition: Duality: The Kantorovich–Rubinstein Theorem

For a ground metric d and distributions $\hat{p}$, $\hat{q}$ on a finite set $\mathcal{X} = \Sigma^n$, the Wasserstein-1 distance satisfies the dual representation

$$W_1(\hat{p}, \hat{q}) \;=\; \sup_{f : \mathrm{Lip}(f) \leq 1} \sum_{g \in \mathcal{X}} f(g) \big[\hat{p}(g) - \hat{q}(g) \big],$$

where the supremum is over all 1-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$ (i.e. $|f(g) - f(g')| \leq d(g, g')$ for all $g, g' \in \mathcal{X}$). This dual is a linear program in the $|\mathcal{X}|$ dual variables $(f(g))_{g \in \mathcal{X}}$, and its optimal solution (the Kantorovich potential) provides a direct characterisation of which n-grams drive the inter-corpus distance.

Sketch. This is the classical Kantorovich duality theorem from optimal transport theory (see, e.g., Villani (2009)). The primal problem (Definition 6.12) is a finite-dimensional linear program; its dual, by the LP duality theorem, involves potentials $f(g)$ and $h(g')$ satisfying $f(g) + h(g') \leq d(g, g')$ for all g, g'. The constraint that d is a metric implies $h(g') = -f(g')$ is optimal, reducing the dual to the supremum over 1-Lipschitz f. □

The Kantorovich–Rubinstein dual has a direct interpretive value in the toponymic setting: the optimal f assigns a real-valued score to each bigram g such that adjacent bigrams (in Hamming sense) receive similar scores, and the W_1 distance is the maximum weighted discrepancy between $\hat{p}$ and $\hat{q}$ under any such smooth scoring function. Bigrams with large $|f(g)|$ are the drivers of the earth mover's distance: they are those for which the corpus contrast is concentrated and geometrically isolated (no neighbouring bigram compensates for the surplus or deficit).

Proposition: Properties of W_1 as a Phonotactic Distance

For n-gram distributions over $(\Sigma^n, d_{\text{edit}})$:

(i) W_1 is a proper metric on the set of probability distributions over Σ^n, satisfying non-negativity, identity, symmetry, and the triangle inequality.

(ii) $W_1(\hat{p}, \hat{q}) \leq n$ (the diameter of Σ^n under the Hamming distance), with equality only when the support of $\hat{p}$ and the support of $\hat{q}$ are disjoint and consist of strings that differ in every character.

(iii) $W_1(\hat{p}, \hat{q}) = 0$ iff $\hat{p} = \hat{q}$.

(iv) W_1 *metrises weak convergence*: a sequence of distributions $\hat{p}_N$ converges in W_1 to $\hat{p}$ iff $\hat{p}_N(g) \to \hat{p}(g)$ for every g with d-bounded support.

Remark: Relationship Between W_1 and KL Divergence in the Toponymic Setting

The Wasserstein-1 and KL distances capture complementary aspects of the difference between two phonotactic distributions. $D_{\mathrm{KL}}(\hat{p}\|\hat{q})$ is sensitive to the *multiplicative* ratio $\hat{p}(g)/\hat{q}(g)$ at each n-gram independently, and diverges to infinity whenever $\hat{p}$ assigns positive mass to an n-gram that $\hat{q}$ assigns zero mass (the zero-support problem motivating Laplace smoothing). W_1, by contrast, is sensitive to the *geometric displacement* of probability mass: if $\hat{p}$ moves mass from bigram g to a neighbouring bigram g' with $d(g, g') = 1$, the contribution to W_1 is $|\hat{p}(g) - \hat{q}(g)| \cdot 1$, whereas the contribution to D_{KL} may be much larger if $\hat{q}(g)$ is very small. In the regime where the two distributions are nearly identical except on a small number of high-frequency bigrams (the Romance-internal confusions of Chapter 11), W_1 provides a more robust and geometrically meaningful quantification of the difference.

The primary reason W_1 is not adopted as a principal measure in the present series is computational: for bigrams over $|\Sigma| = 27$, the linear program has $K^2 = 531{,}441$ primal variables ($K = 729$ bigrams), which is solvable in a fraction of a second by the network-simplex algorithm given the special transportation-polytope structure, but the solution procedure requires substantially more code than the closed-form KL computation and is harder to embed in the bootstrap resampling pipeline of Chapter 3. The Wasserstein distance is nonetheless computable for the Iberian prototype using standard LP solvers, and its zone-pair ordering is consistent with the KL ordering in all comparisons involving inter-family contrasts (Basque vs. Romance) and in most intra-family ones. For the cross-series spatial consolidation extension (Chapter 10), where continuous geographic coordinates rather than discrete zone labels will serve as the spatial index, the Fréchet mean in Wasserstein space (the distribution minimising the expected squared W_2 distance to a sample of n-gram distributions) provides a natural analogue of the zone centroid used in k-means, and its computation via Sinkhorn iterations (Cuturi, 2013) is an established technique that warrants integration into the future pipeline.

Definition: Wasserstein-2 Distance and Fréchet Means

The *Wasserstein-2 distance* uses the squared ground metric d^2 in place of d:

$$W_2(\hat{p}, \hat{q})^2 \;=\; \inf_{\gamma \in \Gamma(\hat{p}, \hat{q})} \sum_{g, g'} d(g, g')^2 \, \gamma(g, g').$$

The *Fréchet mean* (Wasserstein barycenter) of a collection $\{\hat{p}_1, \ldots, \hat{p}_m\}$ with weights $\{\pi_z\}$ is the distribution $\bar{p}$ minimising $\sum_z \pi_z W_2(\hat{p}_z, \bar{p})^2$. For discrete dis-

tributions over Σ^n with the Hamming ground metric, this minimisation is an iterative fixed-point problem solvable by the Sinkhorn–Knopp algorithm, which alternately normalises a coupling matrix by row and column until convergence.

Remark: Why Optimal Transport Is Deferred Rather Than Adopted

The introduction of W_1 and W_2 in this section serves a dual purpose. First, it positions the KL-divergence-based framework of the present series within the broader landscape of distributional comparison methods, showing that the choice of KL over Wasserstein is a deliberate trade-off (interpretability and closed-form computability over geometric sensitivity) rather than a failure to consider alternatives. Second, it provides a concrete research agenda for subsequent work: the open question of whether the zone-pair orderings implied by W_1 and by the symmetric KL divergence agree in the more complex, geometrically richer settings of Vols. V and VI (where the n-gram space exhibits strong geometric structure induced by the consonant-vowel alternation patterns of Dravidian and Armenian) is a well-posed and tractable problem that the data and pipeline of this series are positioned to address.

Chapter 7

Spatial Autocorrelation and Geographic Mutual Information

The cross-regional analyses of Chapter 11 treat each administrative zone as an unstructured unit: phonotactic entropy and KL divergence are computed per zone, but the *geographic arrangement* of zones — which zones are adjacent, how far apart their centroids lie, whether phonotactic similarity decays smoothly with distance or breaks sharply at a linguistic frontier — is not itself quantified. This chapter develops the spatial-statistical apparatus needed to do so, and in particular supplies the formal content of the *spatial mutual information framework* invoked informally in Section 6.5 of Chapter 6.

7.1 Spatial Autocorrelation of Phonotactic Entropy

The starting question is whether phonotactic entropy itself is spatially autocorrelated: do geographically proximate zones have more similar entropy than geographically distant ones, independently of any substrate hypothesis? This is a necessary precondition for any finer-grained spatial substrate mapping (Section 6.5): if entropy varies smoothly and continuously across space, a zone-level partition is merely a convenient discretisation of an underlying continuous field; if entropy is spatially *discontinuous* at certain boundaries, those boundaries are themselves evidence for a genuine substrate frontier rather than an administrative artefact.

Definition: Global Moran's I for a Zone-Level Statistic

Let $\{z_i\}_{i=1}^{m}$ be a statistic (e.g. H_2, or a single n-gram frequency) computed on m zones with centroid coordinates $\{\mathbf{x}_i\}$, and let $W = [w_{ij}]$ be an $m \times m$ spatial weights matrix (e.g. $w_{ij} = 1$ if zones i, j share an administrative border, 0 otherwise, with $w_{ii} = 0$). *Global Moran's I* is

$$I = \frac{m}{S_0} \frac{\sum_i \sum_j w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2}, \qquad S_0 = \sum_i \sum_j w_{ij}.$$

Under the null hypothesis of spatial randomness, $\mathbb{E}[I] = -1/(m-1)$ and $\mathrm{Var}(I)$ has a known closed form under either a normality or a randomisation assumption (Moran, 1950; Cliff and Ord, 1981); a permutation test (randomly reassigning the

$\{z_i\}$ to zones while holding W fixed) is the more robust alternative used throughout this chapter, for consistency with the permutation framework of Chapter 3.

Values of I near $+1$ indicate strong positive spatial autocorrelation (neighbouring zones similar); values near -1 indicate a checkerboard pattern (neighbouring zones dissimilar); $I \approx \mathbb{E}[I]$ indicates spatial randomness.

Proposition: Moran's I for Bigram Entropy, British Isles

Applying the adjacency-based weights matrix to the five British Isles zones (Vol. IV) and the entropy values of Table 11.1 gives $I = 0.31$ under the row-standardised adjacency weighting, against a permutation null of $\mathbb{E}[I] = -0.25$ and a permutation 95% interval of $[-0.51, 0.46]$ (2,000 permutations). The observed value falls within the permutation interval ($p = 0.31$, two-sided), so the entropy field over the five British Isles zones does not depart significantly from spatial randomness at this sample size — an expected and unsurprising result given that $m = 5$ zones affords very limited power for a global autocorrelation test (cf. the power analysis of §7.1.1 below).

Methodological Transparency: Statistical Power of Moran's I at Small m

Every regional volume in this series has $m \in \{4, 5, 5, 6, 6\}$ zones. Global Moran's I has notoriously low power at such small m: the permutation distribution of I under spatial randomness has a standard deviation comparable to its full range, so even pronounced visual clustering in a choropleth map need not register as a significant departure from randomness. The zone-level Moran's I results in this chapter should therefore be read as *descriptive* rather than *confirmatory* — they quantify the degree of spatial smoothness present in the data, but a null result at $m \leq 6$ does not license the inference that no spatial structure exists. Local indicators (Section 7.2) and the continuous kernel-density approach of Section 6.5 are less sensitive to this small-m problem because they do not collapse the spatial information into a single global statistic.

7.1.1 Statistical Power of the Global Test

Because m is fixed by the number of administrative zones in each regional volume, no increase in within-zone corpus size N can improve the power of the global Moran's I test — this is a between-zone, not a within-zone, source of statistical uncertainty, and it is a structural limitation of the zone-level design rather than a data-volume problem. A back-of-envelope power calculation illustrates the point: under a simple autoregressive alternative $z_i = \rho \sum_j w_{ij} z_j / d_i + \eta_i$ with row-degree d_i and i.i.d. noise η_i, the asymptotic variance of I under the null scales as $O(1/m^2)$ for fixed average degree, so the z-statistic for I scales as $O(\sqrt{m}\,\rho)$: detecting a moderate autocorrelation $\rho \approx 0.3$ at conventional power requires $m \gtrsim 25$–30 zones, roughly five to six times the zone count of any single regional volume in the series. This is the single clearest methodological argument, internal to the spatial analysis itself, for the cross-series consolidation now made possible by the completion of Vol. VII, which brings the full zone count to $m = 5 + 6 + 5 + 6 + 6 + 5 = 33$ across Vols. II–VII, finally placing the

global Moran's I test for phonotactic entropy in an adequately powered regime.

7.2 Local Indicators of Spatial Association (LISA)

Global Moran's I collapses spatial structure into a single scalar and, as just discussed, is underpowered at the zone counts available here. The *local* Moran statistic disaggregates the same information to the level of individual zones, and remains informative even when the global test is null, because it does not require averaging away zone-specific signal.

Definition: Local Moran's I_i

The *local Moran statistic* for zone i is

$$I_i = \frac{(z_i - \bar{z})}{\sigma_z^2/(m-1)} \sum_j w_{ij}(z_j - \bar{z}),$$

where $\sigma_z^2 = \frac{1}{m}\sum_i (z_i - \bar{z})^2$. The global statistic decomposes as $I \propto \sum_i I_i$. Each I_i can be classified into one of four quadrants of the Moran scatterplot (zone value vs. spatially lagged neighbour mean): High-High (a high-entropy zone surrounded by high-entropy neighbours), Low-Low, High-Low (a spatial outlier: high entropy surrounded by low), and Low-High.

Example: LISA Classification, Anatolia–Caucasus

Applying the local statistic to the six Anatolia–Caucasus zones with entropy values from Table 11.1 (7.962 Anatolian, 8.141 Greek-Byzantine, 7.894 Caucasian-TR, 7.684 Georgian, 8.174 Azerbaijani, 7.540 Armenian) and a contiguity-based W in which Armenian borders Azerbaijani and Georgian, Azerbaijani borders Armenian and Georgian, and the three Turkic/Anatolian-adjacent zones form a mutually contiguous block: the Turkic block classifies High-High relative to the regional mean only weakly (their entropies are close to the six-zone mean of 7.90), while Armenian classifies unambiguously Low-Low (lowest entropy, bordered by two of the three next-lowest zones) and Azerbaijani classifies High-Low — the highest-entropy zone in the region, but bordering the two lowest (Armenian and Georgian). This Azerbaijani High-Low classification is the spatial-statistical restatement, at the level of a single summary number, of the same KL-divergence pattern already visible in Figure 11.11: Azerbaijani is the most internally diverse neighbour of two markedly more internally coherent zones.

7.3 Spatial Mutual Information

A complementary, information-theoretic formalisation of "does location predict phonotactic structure" replaces the spatial-weights formalism above with a direct mutual-information statistic between a toponym's geographic location and its phonotactic content.

Definition: Spatial Mutual Information

Let $\mathbf{X}$ denote the (discretised, e.g. into a grid of cells) geographic location of a toponym and G the bigram identity of a randomly selected position within it. The *spatial mutual information* of the corpus is

$$I(G;\mathbf{X}) \;=\; \sum_{g}\sum_{\mathbf{x}} p(g,\mathbf{x}) \log_2 \frac{p(g,\mathbf{x})}{p(g)\,p(\mathbf{x})} \;=\; H(G) - H(G \mid \mathbf{X}),$$

the reduction in bigram-identity uncertainty achieved by knowing the toponym's grid cell. $I(G;\mathbf{X}) = 0$ iff G and $\mathbf{X}$ are independent, i.e. iff bigram frequencies are spatially uniform.

This is the natural information-theoretic generalisation of the zone-level KL divergence framework of Chapter 2: the zone partition used throughout this volume is a discretisation $\mathbf{X} \to Z$ (zone label) coarser than the underlying continuous $\mathbf{X}$, and a standard data-processing argument gives $I(G;Z) \le I(G;\mathbf{X})$ — the zone-level analysis can only *lose* spatial information relative to the continuous analysis, never gain it. The gap $I(G;\mathbf{X}) - I(G;Z)$ quantifies how much of the total spatial structure in the corpus is captured by the chosen zone partition versus residing at finer spatial scales within zones (e.g. a smooth gradient from coast to interior that a single provincial zone label cannot represent). Estimating $I(G;\mathbf{X})$ at fine grid resolution requires substantially larger sample sizes than $I(G;Z)$, since the number of effective cells in $\mathbf{X}$ grows with grid resolution while N is fixed; in practice a plug-in estimate with Miller–Madow-style correction for the joint, marginal, and conditional entropies (cf. Chapter 3) is adequate provided the grid is coarse enough that the average cell contains at least $\sim$20–30 toponyms.

7.3.1 Variogram Analysis of Pairwise KL Divergence

A second, purely descriptive diagnostic plots pairwise zone-level KL divergence against the great-circle distance between zone centroids — a phonotactic analogue of the classical semivariogram of geostatistics (Cressie, 1993).

Definition: Phonotactic Semivariogram

For a region with zone centroids $\{\mathbf{x}_i\}$ and pairwise symmetrised divergence $D_{\mathrm{KL,sym}}(i,j) = \tfrac{1}{2}[D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j) + D_{\mathrm{KL}}(\hat{p}_j \| \hat{p}_i)]$, the *empirical phonotactic semivariogram* at lag h is

$$\hat{\gamma}(h) \;=\; \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} D_{\mathrm{KL,sym}}(i,j)^2,$$

where $N(h)$ is the set of zone pairs whose centroid distance falls in a bin around h. A monotonically increasing $\hat{\gamma}(h)$ is consistent with smooth phonotactic decay with distance (a "dialect-continuum" pattern); a flat or non-monotonic $\hat{\gamma}(h)$ indicates that geographic distance is not the dominant driver of phonotactic divergence, and that some other factor (administrative language policy, a sharp ethnic frontier, differential settlement history) better explains the observed divergence pattern.

Example: Semivariogram Pattern, Iberian Zones

Using the five Iberian zone centroids (approximate provincial-capital coordinates) and the symmetrised KL values of Table 5.3, the closest zone pair by centroid distance is Castilian–Galician (adjacent, $D_{\mathrm{KL,sym}} = 0.71$) and the most distant by centroid distance is Basque–Castilian, yet the *largest* divergence in the entire matrix is the geographically intermediate Basque–Catalan pair ($D_{\mathrm{KL,sym}} = 1.98$), not a geographically extreme pair. The semivariogram for this region is therefore markedly non-monotonic, which is the quantitative signature of the qualitative point made repeatedly throughout the series: Basque isolation is a genealogical/typological phenomenon (the sole non-Indo-European zone in the corpus), not a geographic-distance effect, and a naive distance-decay model of phonotactic similarity would systematically mispredict the Basque contrasts.

7.4 Implications for Substrate Boundary Detection

Taken together, the three diagnostics of this chapter — global Moran's I (with its power caveat), local LISA classification, and the semivariogram — supply a principled criterion for distinguishing two qualitatively different kinds of substrate boundary: boundaries that are *continuous* (entropy and KL divergence vary smoothly with distance, consistent with gradual dialectal admixture across a contact zone) and boundaries that are *discontinuous* (a sharp jump in divergence over a short geographic distance, consistent with a genuine typological frontier, e.g. language-family boundary). The Basque case examined above is the clearest instance of the latter in the series to date; the British Isles five-zone Celtic cluster, with its uniformly low pairwise KL values and weak LISA classification, is the clearest instance of the former. Volumes III and V (Central Europe, South Asia) present intermediate cases — e.g. the Pakistani/Indian Indo-Aryan continuum of Vol. V exhibits a semivariogram closer to monotonic distance-decay than either the Basque or British Isles patterns — that the present spatial framework is well-suited to characterise quantitatively in future revisions of this volume, once zone-centroid coordinate data for all six regional volumes are consolidated into a single geocoded dataset.

7.5 Choice of Spatial Weights Matrix

Every statistic in this chapter — global Moran's I, the local LISA decomposition, and the spatial mutual information of Section 7.3 — depends on the spatial weights matrix W through which "neighbouring" is operationalised. The adjacency-based weighting used in the worked examples above ($w_{ij} = 1$ for shared administrative border, 0 otherwise) is the simplest of several standard choices, and the sensitivity of the reported statistics to this choice is a legitimate methodological question independent of any particular volume's data.

Definition: Standard Spatial Weights Specifications

For m zones with centroids $\{\mathbf{x}_i\}$, three weights specifications are in common use:

(i) **Binary contiguity**: $w_{ij} = 1$ if zones i, j share a border, 0 otherwise.

(ii) **Inverse-distance**: $w_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^{-\gamma}$ for $i \neq j$, $\gamma > 0$ a decay exponent, $w_{ii} = 0$.

(iii) **K-nearest-neighbour**: $w_{ij} = 1$ if zone j is among the K nearest centroids to zone i (note: this W is generally asymmetric, since "j is among i's K nearest" need not imply the converse).

Any of the three may additionally be *row-standardised*, replacing w_{ij} with $w_{ij} / \sum_{j'} w_{ij'}$ so that each row sums to 1; row-standardisation is the convention used for the British Isles example above.

Proposition: Effect of Row-Standardisation on Moran's I

Row-standardising W does not change the sign of I relative to the unstandardised matrix when all $w_{ij} \geq 0$, but it does change its magnitude: row-standardisation removes the dependence of S_0 (and hence of the scale factor m/S_0) on the heterogeneity of zone-level degree (number of neighbours), so that zones with many neighbours under binary contiguity (e.g. a central zone bordering four others) do not mechanically dominate the numerator of I relative to a peripheral zone with a single neighbour. For irregular administrative geographies — precisely the case for every regional volume in this series, where some zones (e.g. Castilian in Iberia) border several other zones while others (e.g. Galician) border only one or two — row-standardisation is therefore the more interpretively stable choice and is used throughout this chapter.

Remark: Why Binary Contiguity Rather than Inverse-Distance Weighting

This series uses binary (row-standardised) contiguity rather than inverse-distance weighting for two reasons specific to the toponymic substrate-detection setting. First, administrative zone boundaries in this series are themselves drawn, in most volumes, along historically attested linguistic frontiers (e.g. the Basque Autonomous Community boundary, the Welsh national boundary), so contiguity directly encodes the hypothesis-relevant notion of "adjacent to a documented linguistic frontier," which inverse centroid-distance weighting does not: two zones with distant centroids can nonetheless share a long common border (e.g. large, elongated zones), and conversely two zones with close centroids can be separated by a third zone and therefore non-contiguous. Second, inverse-distance weighting requires choosing the decay exponent γ, an additional researcher degree of freedom with no principled value in the absence of an explicit spatial process model for phonotactic diffusion; contiguity weighting has no such free parameter. The cost of this choice is that contiguity weighting is binary and therefore coarser, discarding information about *how much* border two zones share or how far apart their bulk toponym distributions lie — a limitation partially addressed by the continuous kernel-density approach of Section 6.5, which does not discretise space into zones at all.

7.6 Edge Effects and the Modifiable Areal Unit Problem

A further limitation of any zone-level spatial statistic, orthogonal to the choice of weights matrix, is that the zones themselves are an analytical choice (typically inherited from modern administrative boundaries) rather than a property of the underlying phonotactic field. This is a specific instance of the *Modifiable Areal Unit Problem* (MAUP), well known in quantitative geography (Openshaw, 1984).

Definition: The Modifiable Areal Unit Problem

The MAUP refers jointly to two distinct sensitivities of an areal (zone-level) statistic:

(i) **Scale effect**: aggregating the same underlying point data into coarser or finer areal units changes the value of statistics computed on the aggregated units, even though the underlying point pattern is unchanged.

(ii) **Zoning effect**: for a *fixed* number of areal units, different ways of partitioning the same study region into that number of units yield different values of the aggregated statistic.

Remark: MAUP in the Toponymic Substrate-Detection Setting

Both effects are present, in principle, in every regional volume of this series. The scale effect bears on the zone-count choice itself ($k = 4$–6 administrative zones per volume): a finer partition (e.g. splitting Castilian into separate northern and southern sub-zones) would likely reveal additional within-zone heterogeneity at the cost of smaller per-zone corpus sizes and reduced estimation precision (Chapter 3), illustrating a direct trade-off between MAUP scale sensitivity and statistical power that this series resolves in favour of fewer, larger, better-estimated zones. The zoning effect bears on the specific choice of administrative boundaries used to define each zone (Section 5.4.6); the sensitivity analysis already reported there for zone-boundary re-definition is, in MAUP terminology, a direct empirical probe of the zoning effect, conducted at fixed zone count k.

Proposition: MAUP and the Robustness of Substrate-Level (Rather than Zone-Level) Conclusions

The MAUP affects zone-*level* statistics (entropy per zone, Moran's I over zones) more severely than it affects *pairwise, substrate-identified* statistics such as $D_{\mathrm{KL}}(\widehat{\text{Basque}} \| \widehat{\text{Castilian}})$, because the latter are computed by pooling all toponyms independently assigned to a named substrate category (a linguistic, not administrative, unit) rather than by aggregating within an arbitrary areal partition. To the extent that a substrate zone's boundary in this series approximates the true linguistic-historical extent of that substrate (e.g. the Basque Autonomous Community boundary approximating, though not equalling, the historical extent of Basque toponymy), the zoning effect on the *pairwise* KL divergences reported throughout the series is bounded by the proportion of misassigned toponyms near the zone boundary, a quantity directly estimated by the boundary-

sensitivity analysis of Section 5.4.6 rather than by the MAUP literature's generic bounds.

Chapter 8

Systematic Morphological Comparison of the Six Onomastic Systems

Chapters 11 and 7 compare the six regional volumes purely at the level of aggregate statistics: entropy, KL divergence, perplexity, spatial autocorrelation. This chapter takes a complementary, more linguistically grounded approach, asking what *morphological mechanisms* underlie the observed phonotactic contrasts. Each onomastic system in the series belongs to a different broad morphological type, and the entropy and KL results of the preceding chapters turn out, on inspection, to track these typological differences closely.

8.1 A Typology of Toponymic Morphology

Definition: Four Morphological Strategies in Toponym Formation

Across the six regional corpora, toponym-forming morphology can be classified into four (non-exclusive) strategies:

(i) **Compounding**: concatenation of a generic term (*river*, *mountain*, *settlement*) with a specific qualifier, with or without a linking morpheme;

(ii) **Affixation**: derivation of a place name from a root (often a personal name or common noun) via a productive suffix or prefix paradigm;

(iii) **Agglutination**: stacking of multiple bound morphemes, each contributing a discrete semantic increment, with comparatively transparent morpheme boundaries;

(iv) **Fusion**: morphological material historically derived from (i)–(iii) but no longer synchronically segmentable, having undergone sufficient sound change that the etymological boundaries are recoverable only via historical-comparative reconstruction.

8.2 Iberian Zones: Fusional Romance vs. Agglutinative Basque

The Iberian corpus (Vol. II) juxtaposes the clearest fusional/agglutinative contrast in the series. Castilian, Catalan, and Galician toponyms are overwhelmingly suffixally derived from Latin bases via a small, closed set of productive suffixes (*-illo*/*-ito* diminutives, collective *-eda*/*-edo*, possessive adjectival formations), with extensive subsequent sound change (vowel diphthongisation, lenition of intervocalic consonants) that has fused root and suffix into a largely unsegmentable phonological word. Basque toponym formation, by contrast, is agglutinative in the canonical sense: a toponym such as *Berrioplano* or *Etxaberria* can typically be decomposed into a sequence of transparent bound morphemes (*etxe*-'house', *berri*-'new', *-a* article) whose boundaries remain synchronically recoverable. This is the morphological substrate of the higher Basque bigram entropy reported throughout this volume (Table 5.2): agglutination permits a substantially larger inventory of productive bound morphemes than the closed Romance suffix paradigms, and each additional morpheme type contributes additional bigram diversity at its boundary with adjacent morphemes.

Proposition: Morpheme Inventory and Bigram Entropy

For a toponym-formation system with M productive bound morphemes combining with approximately uniform frequency, the bigram entropy contributed by morpheme-boundary transitions scales, to first order, as $\log_2 M$ (the entropy of a uniform distribution over M boundary types), holding root-internal phonotactics fixed across systems. A system with a larger productive morpheme inventory will therefore exhibit higher *boundary* bigram entropy even if its root-internal phonotactics are equally constrained.

This proposition is consistent with, though it does not by itself fully explain, the observed Basque–Romance entropy gap: Basque toponymic morphology productively combines on the order of several dozen common bound elements (generic terms for 'house', 'new', 'stone', 'oak', etc., plus a handful of grammatical suffixes), whereas the closed Romance diminutive/collective suffix paradigms number in the single digits.

8.3 British Isles: Celtic Mutation Systems and Romance-Anglo-Saxon Layering

The five British Isles zones (Vol. IV) illustrate a different morphological axis: *initial consonant mutation*. Welsh, the source of the highest-entropy zone in that volume (Table 11.1), productively mutates the initial consonant of a qualifying element following certain grammatical triggers (soft mutation $p \to b$, $t \to d$, $c \to g$; nasal mutation; aspirate mutation), so that the *same* lexical root surfaces with phonologically distinct initial bigrams depending on its syntactic position within the toponym. This is the direct morphophonological source of the elevated `wy`, `rh`, and `wr` bigram frequencies documented for Wales in Table C.3: each of these reflects a mutation-conditioned onset that has no productive counterpart in the non-Celtic zones of the same volume.

Scotland and Ireland exhibit a structurally similar (Goidelic) mutation system (lenition realised orthographically via following *h*), which is the source of the `bh`, `mh`, and `ch` marker bigrams shared by those two zones but largely absent from England and much reduced in Wales. England, finally, is morphologically the "odd one out" of the five British Isles zones: its toponymic stock is a diachronic layering of Brittonic-Celtic substrate elements beneath an Anglo-Saxon (Old English) compounding system (*-ton*, *-ham*, *-ley*, *-field* as generic second elements) with a later Norse adstrate in the north and east, none of which involve productive synchronic mutation. The narrow entropy spread among all five zones noted in Figure 11.3's caption is therefore somewhat surprising on purely morphological grounds, since two structurally quite different systems (Celtic mutation vs. Anglo-Saxon compounding) yield similar aggregate bigram entropy; this is consistent with the general observation (Chapter 11) that aggregate entropy is a comparatively coarse summary that can mask substantial differences in the *mechanism* generating a given entropy value, motivating the present chapter's more granular, mechanism-level comparison.

8.4 Anatolia–Caucasus: Agglutinative Turkic vs. Fusional Indo-European vs. Ergative Caucasian

The Anatolia–Caucasus corpus (Vol. VI) contains the typologically richest contrast in the series. Azerbaijani and the Anatolian Turkic layer are agglutinative with productive vowel harmony, a morphophonological process whereby suffix vowels assimilate in backness and rounding to the preceding root vowel; this is plausibly implicated in the comparatively high, evenly distributed bigram frequency profile noted for Anatolian in Chapter C.2 (vowel harmony tends to homogenise the inventory of permissible root-suffix bigram transitions across many different roots, since the same small set of harmonic vowel pairs recurs at every morpheme boundary regardless of which specific suffix is involved, spreading frequency mass relatively evenly across many such transitions rather than concentrating it). Georgian, a fusional/agglutinative language with a complex consonant-cluster system (geologically, "no-vowel" consonant sequences such as the classic *gvbrdgvnis*-type clusters are well attested, though toponym-internal clusters are typically less extreme), contributes a phonotactic profile dominated by consonant-cluster bigrams largely absent from the Turkic zones. Armenian, finally, combines a moderately rich but closed inflectional-suffix paradigm with a comparatively small productive toponym-formation vocabulary relative to the other zones in this volume, which is the most plausible morphological correlate of its markedly lower bigram entropy and its top-heavy bigram frequency distribution documented in Chapter C.2 (a single bigram, `ar`, accounting for 5.58% of all bigram tokens, more than double its frequency in any other zone in the volume) — a smaller productive formation vocabulary concentrates frequency mass onto a correspondingly smaller number of recurrent root and suffix bigrams.

Table 8.1: Qualitative morphological classification of the highest- and lowest-entropy zone in each regional volume (entropy values from Table 11.1), illustrating the association between morphological type and entropy rank discussed in this chapter.

Volume	**Highest-entropy zone**	**Morphological type**	H_2
II (Iberia)	Basque	Agglutinative	7.652
III (Central Europe)	Alpine-Celtic	Celtic, mixed substrate	8.108
IV (British Isles)	Wales	Celtic, mutation system	7.871
V (South Asia)	PK-Plains	Indo-Aryan, compounding	7.558
VI (Anatolia–Caucasus)	Azerbaijani	Agglutinative, vowel harmony	8.174
Volume	**Lowest-entropy zone**	**Morphological type**	H_2
II (Iberia)	Castilian	Fusional Romance	7.318
III (Central Europe)	Etruscan	Fusional, heavily Romanised	7.308
IV (British Isles)	Scotland	Celtic, mutation, reduced	7.624
V (South Asia)	Indo-Aryan-E	Indo-Aryan, fusional	7.057
VI (Anatolia–Caucasus)	Armenian	Fusional, closed paradigm	7.540

8.5 Cross-System Synthesis: Morphological Type as a Predictor of Entropy Rank

The pattern in Table 8.1 is suggestive rather than statistically established (it is a classification of five highest/lowest pairs, far too few observations for any formal test): in four of the five volumes, the highest-entropy zone is agglutinative or exhibits a productive mutation/harmony system, while the lowest-entropy zone is fusional with a comparatively closed formation paradigm. South Asia (Vol. V) is the partial exception, since both the highest- and lowest-entropy zones are broadly Indo-Aryan; there, the operative contrast (PK-Plains' extreme length variance versus Indo-Aryan-East's comparative regularity) appears to be a corpus register/text-source effect compounding the morphological factor more than a clean typological contrast. The broader methodological lesson, consistent with the explicit caveats of Chapter 11, is that aggregate entropy and KL divergence are powerful for *detecting* that two corpora differ, but identifying *why* they differ — the morphological mechanism — requires precisely the kind of qualitative, linguistically informed analysis undertaken in this chapter, and the two levels of analysis should be regarded as complementary rather than substitutable.

8.6 Indian Subcontinent: Indo-Aryan Compounding, Dravidian Suffixation, and Munda Roots

The South Asian corpus (Vol. V) occupies a middle position in the series entropy hierarchy (7.057–7.558 bits), and its morphological landscape is the most typologically heterogeneous in the series, spanning at least three unrelated language families.

Indo-Aryan: compounding with Sanskrit nominal morphology. The highest-entropy zone in the South Asian corpus, PK-Plains, is dominated by Pakistani urban

and rural toponyms of predominantly Indo-Aryan origin (Punjabi, Urdu, Sindhi), supplemented by a substantial Arabic and Persian stratum from the Mughal administrative vocabulary (*-abad*, *-nagar*, *-pur*). This multi-source compound morphology is the primary driver of the relatively high PK-Plains entropy: the toponym-final generic elements (*-abad* 'settlement', *-kot* 'fortress', *-wala* 'belonging to') each introduce a distinct and highly frequent final bigram sequence, and the diversity of these competing generic suffixes increases the effective bigram vocabulary beyond what a single-source suffix paradigm would produce.

The lowest-entropy Indo-Aryan zone (Indo-Aryan-E, covering Bengal, Bihar, and Odisha) has a much more homogeneous morphological profile: Bengali and Odia toponymy is dominated by a narrow set of suffixes derived from Sanskrit *-gram* 'village' (modern *-gram*, *-gaon*), *-para* 'neighbourhood', *-hat* 'market', and *-pur* 'town'. This typological homogeneity — fewer competing suffix types, all ending in the same small set of final consonants — is the morphological correlate of the comparatively low Indo-Aryan-E entropy.

Dravidian: verb-final morphology and postpositional suffixes. The Dravidian zone (covering Tamil Nadu, Kerala, and Karnataka) exhibits a morphological profile that is distinctive in the series: Dravidian toponymy is characterised by heavy use of postpositional case suffixes (*-ur* 'town/settlement', *-patti* 'hamlet', *-kudi* 'hamlet', *-nadu* 'land') that are highly productive and that follow the toponym root with transparent boundaries. The Dravidian pattern is thus structurally agglutinative (morpheme boundaries are recoverable), yet the effective morpheme inventory for toponym-formation is comparatively small: five to eight generic elements account for the vast majority of Dravidian toponymic suffixes. This intermediate morphological profile is consistent with the intermediate entropy position of the Dravidian zone (7.108 bits, Table 11.1).

Munda: poly-morphemic roots and non-Indo-European phonotactics. The Munda zone (covering Jharkhand and adjacent tribal areas) provides the most unusual phonotactic profile in the South Asian corpus. Munda languages (Santali, Ho, Mundari) are polysynthetic at the morphological level but exhibit a highly constrained phoneme inventory by South Asian standards: they lack the retroflex consonant series that dominates Indo-Aryan phonotactics, lack the aspirated stop series that is characteristic of both Indo-Aryan and Dravidian, and have a strong preference for CVCV syllable structures that produce a highly regular alternating consonant-vowel bigram profile. The Munda zone entropy (7.092 bits) is slightly below the Dravidian zone, consistent with the greater regularity of the CVCV template: high bigram entropy requires a diverse inventory of consonant-vowel co-occurrence patterns, and the restricted Munda phoneme inventory combined with the regular syllabic template limits this diversity.

Proposition: Syllabic Template and Bigram Entropy

For a corpus $\mathcal{T}$ generated by a strictly CVCV syllabic template (consonant $\in \mathcal{C}$, vowel $\in \mathcal{V}$, with $|\mathcal{C}| = n_C$ and $|\mathcal{V}| = n_V$), the maximum achievable bigram entropy

is

$$H_2^{\max}(\text{CVCV}) \;=\; \tfrac{1}{2}\log_2 n_C + \tfrac{1}{2}\log_2 n_V + \tfrac{1}{2}\log_2 n_V + \tfrac{1}{2}\log_2 n_C \;=\; \log_2(n_C n_V),$$

since the CVCV template enforces that a consonant bigram is always VC or CV (not CC or VV), and the maximum entropy is achieved when all consonant-vowel and vowel-consonant transitions are equiprobable. For Munda languages with $n_C \approx 18$ and $n_V \approx 6$, this gives $\log_2(108) \approx 6.75$ bits — a ceiling lower than the observed 7.09 bits (which is elevated above the pure CVCV prediction by the presence of occasional consonant clusters and word-boundary sequences), consistent with the zone being near but not at the CVCV entropy ceiling.

Remark: Confounding by Orthographic Convention

The South Asian corpus presents a distinctive orthographic challenge: GeoNames entries for Indian toponyms draw on multiple romanisation traditions (ITRANS, National Library at Calcutta, ISO 15919, and various government-standard transliterations) that assign different Latin-script bigrams to the same phonological sequence. For example, the retroflex lateral approximant of Dravidian languages is romanised as `l` in standard Tamil romanisation, as `ll` in colloquial orthography, and as `l.` in the ISO standard. The pipeline normalises these to the most common form per GeoNames source file, but residual variance across romanisation schemes contributes an estimated 0.03–0.05 bits to within-zone entropy for the Dravidian and Munda zones.

8.7 Central Europe: Extinct Substrates and the Depth of Romanisation

The Central European corpus (Vol. III) has the widest entropy range in the series (0.800 bits, Table 11.1), driven by the contrast between Alpine-Celtic (the highest-entropy zone in the series at 8.108 bits) and Etruscan (the lowest at 7.308 bits within Vol. III). The morphological mechanisms behind this contrast deserve explicit analysis.

Alpine-Celtic: a large consonant inventory and compounding. Gaulish, the Celtic language responsible for the Alpine-Celtic substrate, was a P-Celtic language with a large consonant inventory relative to Latin: it retained labiovelars (qu-/p-) as separate phonemes, had a rich series of spirants, and employed extensive compounding of generic geographic terms (**brig-* 'hill', **nemeto-* 'sacred grove', **dun-* 'fort') with determinative qualifiers. The compound structure generates high-entropy bigram profiles because (a) the consonant clusters at morpheme boundaries (e.g. *-brig-* contributing `br` and `ig`) are diverse relative to Latin-derived morphology, and (b) the compounding is productive rather than closed: the number of attested Gaulish geographic generics is large relative to the number of productive Romance suffix types, yielding a correspondingly larger effective bigram vocabulary.

Etruscan: phonological filtering through Latin. Etruscan toponyms in central Italy have been transmitted almost entirely through Latin administrative records, and the process of Latin transmission involved systematic phonological adaptation: Etruscan voiced stops were typically adapted to Latin voiced stops, Etruscan fricatives were approximated by Latin sonorants or dropped, and the Etruscan characteristic final-vowel aspiration left no trace in the Latin alphabet. The surviving Etruscan-substrate toponyms in modern Italian — names like *Volterra*, *Cortona*, *Vetulonia* — have been so thoroughly filtered through Latin and then through medieval Tuscan phonology that their Etruscan phonotactic signature is heavily diluted. This is the morphological and phonological reason why the Etruscan zone has the lowest entropy in Vol. III despite Etruscan being a genuine typological isolate: the substrate signal has been eroded by multiple layers of linguistic filtering in a way that has no parallel for the Alpine-Celtic zone (which was never as deeply Romanised as central Italy).

Example: Romanisation Depth as an Entropy Predictor

A proxy for Romanisation depth is the proportion of toponyms in a zone whose final syllable is one of the Latin productive suffixes (*-us*, *-a*, *-um*, *-is*, *-ensis*). For the Etruscan zone this proportion is approximately 68%; for the Alpine-Celtic zone it is approximately 41%. The difference is consistent with the 0.800 bit entropy gap between the zones: more Latinised toponyms have their substrate-specific bigrams replaced by the narrow inventory of Latin suffixal bigrams (`us`, `is`, `um`), compressing the bigram distribution toward the Latin entropy baseline.

8.8 Etymology-Based Substrate Assignment

The traditional method of substrate assignment — individual etymological analysis — provides a natural benchmark for the present framework. In principle, if every toponym in a corpus could be reliably assigned to a specific substrate layer through etymology, the ground truth would be known and the information-theoretic approach would be unnecessary.

In practice, etymology-based assignment faces three fundamental limitations that the present framework avoids.

Sparsity. For most regions, only a minority of toponyms have been assigned to specific substrate layers in the scholarly literature. For Iberia, the most thoroughly studied case, perhaps 2–3% of all GeoNames-listed toponyms have individually published etymologies; for Central Europe and the British Isles, the proportion is lower. The information-theoretic framework uses the full corpus including unassigned toponyms and draws strength from the aggregate distributional signal.

Circularity. Individual etymology is vulnerable to confirmation bias: the researcher's prior belief about the substrate configuration influences the assignment. Quantitative validation of etymological hypotheses typically uses the same etymological reasoning that generated them. The information-theoretic framework is entirely independent of individual etymologies; its results can therefore be used as a

non-circular check on etymological hypotheses.

Language-dependence. Etymological analysis requires deep expertise in each attested and reconstructed language involved. For regions with extinct, poorly-attested substrates (Etruscan, Messapic, Iberian, Urartian), the expertise base is narrow. The information-theoretic framework requires only a clean corpus and a specification of geographic proxy zones; no linguistic knowledge beyond basic normalisation is needed for the computational analysis.

8.9 Typological Distance Measures

Typological databases such as WALS (Dryer and Haspelmath, 2013) and ASJP (Wichmann et al., 2010) provide phonological distance measures between language pairs. These measures are based on the typological features or basic vocabulary of modern (or historically attested) languages, not on their substrate traces in toponymy.

The present framework differs from typological distance in three respects. First, it operates on *substrate traces*, not on the languages themselves; a language may leave a substrate trace that differs from its synchronic phonology due to the filtering effects of toponym assimilation. Second, it does not require the substrate language to be attested (the Basque signal would be recoverable even if Basque had become extinct). Third, it is geographically anchored to specific regions, making it sensitive to regional substrate mixing in a way that language-level measures are not.

8.10 Automated Language Identification

Character n-gram language identification systems (Cavnar and Trenkle, 1994) use the same feature representation (n-gram profiles) as the present framework, but in the reverse direction: given a known language model and an unknown text, they classify the text. Applied to toponymy, such systems could be used to classify individual toponyms by substrate.

The key difference is that language identification systems require training data from the target language(s), whereas the present framework requires only unlabelled geographic data. Where historical languages are extinct and no corpus of their toponymy exists, language identification is inapplicable; the information-theoretic framework remains applicable as long as a geographic proxy zone can be defined.

Chapter 9

Sensitivity Analysis and Robustness

9.1 Sensitivity to Corpus Size

The most significant parameter choice in the pipeline is the corpus size N per zone. Too small a corpus produces unreliable entropy estimates; too large a corpus may introduce geographic inhomogeneity within zones.

Using the Iberian prototype as a reference, the pipeline was applied at corpus sizes $N \in \{200, 500, 1{,}000, 2{,}000, 3{,}000, 5{,}000\}$ per zone, with results reported in Table 9.1.

Table 9.1: Sensitivity of principal results to corpus size N per zone, Iberian prototype. Basque isolation ratio = $\min_j D_{\mathrm{KL}}(\mathrm{Ba}\|\mathrm{j}) / \max_{j,\ell\neq\mathrm{Ba}} D_{\mathrm{KL}}(\hat{p}_j\|\hat{p}_\ell)$.

N	H_2(Basque) (bits)	Basque isolation ratio	Mean purity
200	7.31	1.07	0.27
500	7.46	1.14	0.29
1,000	7.58	1.22	0.31
1,896	7.65	1.28	0.32
(*Basque pool exhausted at $N = 1{,}896$; larger N uses all available Basque toponyms*)			

The results are stable for $N \geq 1{,}000$. The Basque pool in the real GeoNames corpus is exhausted at $N = 1{,}896$; all analyses therefore use all available Basque toponyms, with bootstrapped resampling used to assess variance.

Expanding the zone decreases the Basque isolation ratio from 1.278 to approximately 1.19 (the additional toponyms from Romanised Navarre dilute the Basque signal). Contracting the zone to the three provinces of the Basque Autonomous Community increases it to approximately 1.34 (the purest Basque area). The qualitative conclusion — that the Basque zone satisfies the Substrate Isolation Criterion — is robust to all tested boundary variants.

9.2 Sensitivity to Normalisation Parameters

Stoplist size. Results are remarkably insensitive to the size of the generic stoplist, provided the list does not include terms that are genuinely toponymic (cf. Section 4).

Expanding the stoplist from 200 to 500 terms per language changes the bigram entropy of no zone by more than 0.015 bits.

Minimum length filter. Changing the minimum toponym length from 4 to 3 or 5 characters affects the entropy estimates by at most 0.02 bits, well within the bootstrap confidence intervals. The mean purity of clustering is unaffected ($\pm 0.5\%$).

Smoothing parameter. The KL divergence is sensitive to the smoothing parameter ε when one distribution assigns near-zero probability to an n-gram that the other assigns moderate probability to. With $\varepsilon = 10^{-9}$ (near-zero smoothing), the divergences reported in Table 5.3 are obtained. With $\varepsilon = 1$ (Laplace smoothing), all divergences decrease by approximately 15–20%, but the ordering and the magnitude of the Basque–Romance contrast are preserved.

9.2.1 Compound Sensitivity: Combined Variations

Beyond single-parameter perturbations, it is instructive to examine whether combined changes to several normalisation parameters simultaneously drive the conclusions away from those of the default pipeline. Table 9.2 summarises a factorial experiment varying three binary parameters — stoplist size (small 200 / large 500), minimum toponym length (3 / 5 characters), and smoothing (near-zero 10^{-9} / Laplace 1) — across all $2^3 = 8$ combinations, with the Basque isolation ratio as the outcome variable.

Table 9.2: Compound sensitivity: Basque isolation ratio $\min_j D_{\mathrm{KL}}(\mathrm{Ba} \| \hat{p}_j) / \max_{j,\ell \neq \mathrm{Ba}} D_{\mathrm{KL}}(\hat{p}_j \| \hat{p}_\ell)$ under all combinations of three normalisation parameters. Default configuration shaded; all variants satisfy the Criterion at $\delta > 1$.

Stoplist	Min length	Smoothing	Isolation ratio	SIC met?
200	3	10^{-9}	1.24	yes
200	3	1	1.11	yes
200	5	10^{-9}	1.29	yes
200	5	1	1.13	yes
⋆	4	10^{-9}	1.278 (default)	yes
500	3	10^{-9}	1.22	yes
500	3	1	1.09	yes
500	5	10^{-9}	1.27	yes
500	5	1	1.12	yes

Across all eight configurations, the Basque isolation ratio remains above 1.07, and the Substrate Isolation Criterion is satisfied in every case. The minimum ratio (1.07, under maximal stoplist expansion, minimum length 3, and Laplace smoothing) represents the worst-case combination: the large stoplist removes the most surface material from toponyms; the shorter length filter admits many 3-character strings whose bigram profile is dominated by boundary-character transitions (#a, a#) that are identical across all zones; Laplace smoothing compresses the KL divergences uniformly.

Even under this combination, the isolation ratio remains above 1, confirming the robustness of the Basque signal to joint adverse parameter choices.

9.2.2 Sensitivity to Vocabulary Cutoff

The vocabulary is restricted to the top-K most frequent n-gram types (default $K = 500$ for bigrams, $K = 2{,}000$ for trigrams). Table 9.3 reports the effect of varying K on the principal Iberian results.

Table 9.3: Effect of bigram vocabulary cutoff K on entropy and KL divergence estimates, Iberian prototype. Entropy values are for the Basque zone; KL values are $D_{\mathrm{KL}}(\mathrm{Ba}\|\mathrm{Ca})$ as a representative large-divergence pair.

K	$H_2(\mathrm{Ba})$	$H_2(\mathrm{Ca})$	$D_{\mathrm{KL}}(\mathrm{Ba}\|\mathrm{Ca})$	Isolation ratio
100	7.091	6.832	2.56	1.49
200	7.348	7.117	2.81	1.35
500	7.652	7.318	0.934	1.278
1000	7.694	7.352	0.948	1.263
2000	7.705	7.361	0.951	1.258

Entropy and KL values increase monotonically with K, converging by $K \approx 1{,}000$. The isolation ratio decreases from 1.51 at $K = 100$ to 1.23 at $K = 2{,}000$, reflecting the fact that rarer n-grams tend to be less zone-specific (they occur at low frequency in multiple zones) and therefore dilute the isolation signal when included. The default $K = 500$ represents a conservative choice that includes all n-gram types with frequency above approximately 0.1% of total tokens and is positioned on the stable plateau of the convergence curve.

9.3 Jackknife Variance Estimation

The bootstrap procedure of Chapter 3 resamples toponyms *with* replacement; the delete-d jackknife provides an independent variance estimate by systematically deleting subsets and is useful as a cross-check, since the two procedures make different asymptotic assumptions and can disagree when the underlying n-gram count distribution is sufficiently heavy-tailed.

Definition: Delete-d Jackknife Entropy Variance

Partition $\mathcal{T}$ ($|\mathcal{T}| = N$) into $k = N/d$ disjoint blocks of size d. For each block b, compute $\hat{H}^{(-b)}$, the Miller–Madow entropy estimate with block b removed. The jackknife variance estimate is

$$\widehat{\mathrm{Var}}_{\mathrm{jack}}(\hat{H}) = \frac{k-1}{k}\sum_{b=1}^{k}\left(\hat{H}^{(-b)} - \bar{H}^{(\cdot)}\right)^2, \qquad \bar{H}^{(\cdot)} = \frac{1}{k}\sum_{b}\hat{H}^{(-b)}.$$

Applied to the Iberian prototype with $d = N/20$ (20 jackknife blocks), the jackknife standard error for Basque bigram entropy is 0.0041 bits, against a bootstrap

standard error of 0.0039 bits from the same corpus — agreement to within 5%, which is the expected magnitude of difference between the two estimators at this sample size and provides an internal consistency check on the bootstrap procedure used as the primary uncertainty quantification throughout this series. A jackknife/bootstrap discrepancy substantially larger than this (say, a factor of two or more) would indicate that the n-gram count distribution is sufficiently heavy-tailed that the bootstrap's implicit i.i.d. resampling assumption is being stressed, and would warrant a more conservative (e.g. block-bootstrap) variance estimate; no such discrepancy is observed for any zone in the series.

9.4 Cross-Validation via Leave-One-Zone-Out

A final, qualitatively different robustness check asks not whether the entropy or KL *estimates* are stable, but whether the substantive *conclusions* of the series (e.g. "Basque satisfies the Substrate Isolation Criterion") are sensitive to the inclusion of any single zone in the comparison set.

Definition: Leave-One-Zone-Out Re-Ranking

For a region with m zones, the *leave-one-zone-out* procedure removes each zone i in turn, recomputes the isolation ratio of the Substrate Isolation Criterion (Theorem 2.8, Chapter 2) for every remaining zone using only the $m-1$ zones that remain, and checks whether the qualitative ranking of zones by isolation ratio is preserved.

Example: Leave-One-Zone-Out, Iberian Prototype

Removing Arabic from the five-zone Iberian comparison and recomputing the isolation ratio for the remaining four zones leaves Basque's rank unchanged (still the unique maximum, with an even larger margin over the runner-up, since Arabic was itself the second-most distinctive zone after Basque). Removing Catalan, Galician, or Castilian in turn similarly leaves Basque's top rank unchanged in all three cases, with isolation-ratio margins fluctuating by at most 8%. The substantive conclusion of Volume II — that Basque, and only Basque, satisfies the Substrate Isolation Criterion among the five Iberian zones — is therefore robust to the removal of any single comparison zone, which is the relevant robustness property for a claim that is inherently *relative* to a comparison set: a substrate isolation claim that depended sensitively on which *other* zones happened to be included in the comparison would be considerably weaker evidence for a genuine substrate effect than one which, as here, is stable under leave-one-out perturbation of the comparison set itself.

9.4.1 Full Leave-One-Zone-Out Results

Table 9.4 reports the full leave-one-zone-out results for all five Iberian zones: in each column, one zone is removed and the isolation ratios for the remaining four zones are recomputed. The cell in position (i, j) gives the isolation ratio of zone i when zone

j has been removed. Diagonal cells (zone removed = zone assessed) are undefined and left blank.

Table 9.4: Leave-one-zone-out isolation ratios for the Iberian prototype. Each column removes one zone; each row reports the isolation ratio of that zone in the reduced comparison. Cells with ratio > 1 (SIC satisfied) are in plain type; ratios ≤ 1 would appear in italics (none observed). Ba = Basque, Ar = Arabic, Ca = Castilian, Cat = Catalan, Ga = Galician.

Zone assessed	**Zone removed**				
	Ba	Ar	Ca	Cat	Ga
Basque	—	1.38	1.25	1.31	1.27
Arabic	1.06	—	1.09	1.05	1.08
Castilian	1.04	1.07	—	1.02	1.03
Catalan	1.11	1.09	1.14	—	1.10
Galician	1.09	1.11	1.10	1.08	—

Several observations follow from Table 9.4:

(i) Basque satisfies the SIC in every leave-one-out configuration with ratios ranging from 1.25 (removing Castilian) to 1.38 (removing Arabic, the second most-distinctive zone). The Basque isolation is not an artefact of any single comparison zone.

(ii) Arabic satisfies the SIC in every configuration, but with notably lower ratios (1.05–1.09), consistent with its position as phonotactically distinctive from the Romance zones while nonetheless belonging to the same Semitic-influenced Romance contact zone. When Basque is removed, Arabic's ratio rises from 1.06 to 1.09 (the denominator of the SIC formula decreases because Basque was the widest pairwise gap among non-Arabic zones).

(iii) The three Romance zones (Castilian, Catalan, Galician) each satisfy the SIC, but with ratios close to 1 (1.02–1.14), indicating that their isolation from one another is the most fragile in the corpus and could in principle disappear if the corpus were smaller or the zone boundaries redefined to capture more admixed transitional areas.

9.5 Non-Parametric Permutation Bounds on Estimation Error

Beyond the bootstrap and jackknife, a complementary view on estimation uncertainty uses a fully non-parametric permutation argument. The key question is: if the true n-gram distributions of all zones were identical (the null of no substrate structure), how large could the observed pairwise KL divergences be by chance at the given corpus sizes?

Proposition: Permutation Bound on Spurious KL

Under the null hypothesis that all zones are drawn from a single common distribution p, the expected value of the plug-in KL divergence $\widehat{D_{\mathrm{KL}}}(\hat{p}_i \| \hat{p}_j)$ over random zone assignments is

$$\mathbb{E}_{\text{perm}}\left[\widehat{D_{\mathrm{KL}}}(\hat{p}_i \| \hat{p}_j)\right] \;\approx\; \frac{|\Sigma^n| - 1}{2 \ln 2} \left(\frac{1}{N_i} + \frac{1}{N_j} \right)$$

to first order in N^{-1}, where $|\Sigma^n|$ is the vocabulary size. For bigrams over a 27-character alphabet ($|\Sigma^2| = 729$), $N_i = N_j = 3{,}000 \times 8 = 24{,}000$ tokens, this evaluates to approximately 0.0088 bits.

All pairwise KL divergences among Iberian zones (Table 5.3) are at least an order of magnitude above this permutation-null expectation: the smallest non-zero entry is 0.089 bits ($D_{\mathrm{KL}}(\mathrm{Ca} \| \mathrm{Ar})$), a factor of ten above 0.0088 bits. The Basque entries are several hundred times the null expectation. This confirms that the observed divergences cannot plausibly arise from chance sampling variation, independently of the formal Bayes factor computation of Section 5.

9.6 Asymptotic Distribution of the Plug-in KL Estimator

The permutation bound of the previous section was stated without derivation. This section derives it from first principles via the delta method, which also clarifies the regime of validity of the approximation and its relationship to the bias-correction results of Chapter 3.

Proposition: Null-Case Bias of the Plug-in KL Estimator

Let $\hat{p}_i$ and $\hat{p}_j$ be plug-in n-gram distribution estimates from two independent samples of size N_i and N_j, both drawn from the *same* true distribution p (the null of no substrate difference). Then, to leading order in N_i^{-1} and N_j^{-1},

$$\mathbb{E}\left[\widehat{D_{\mathrm{KL}}}(\hat{p}_i \| \hat{p}_j)\right] \;\approx\; \frac{|\Sigma^n| - 1}{2 \ln 2} \left(\frac{1}{N_i} + \frac{1}{N_j} \right).$$

Proof. Write $\hat{p}_i(g) = p(g) + \delta_i(g)$ and $\hat{p}_j(g) = p(g) + \delta_j(g)$, where δ_i, δ_j are zero-mean sampling fluctuations with $\mathrm{Var}(\delta_i(g)) = p(g)(1-p(g))/N_i$ (the multinomial variance of a single cell proportion; cf. Section 3.1) and $\delta_i \perp \delta_j$ by independence of the two samples. A second-order Taylor expansion of $\widehat{D_{\mathrm{KL}}}(\hat{p}_i \| \hat{p}_j) = \sum_g \hat{p}_i(g) \log_2[\hat{p}_i(g)/\hat{p}_j(g)]$ around $\delta_i = \delta_j = 0$ gives, after the first-order terms vanish in expectation (since $\mathbb{E}[\delta_i(g)] = \mathbb{E}[\delta_j(g)] = 0$),

$$\widehat{D_{\mathrm{KL}}}(\hat{p}_i \| \hat{p}_j) \;\approx\; \frac{1}{2 \ln 2} \sum_g \frac{[\delta_i(g) - \delta_j(g)]^2}{p(g)} + O(N^{-3/2}),$$

using the standard second-order expansion of $x \log(x/y)$ around $x = y = p(g)$ (the same expansion underlying the Miller–Madow correction of Section 3.2.1). Taking

expectations and using independence of δ_i and δ_j,

$$\mathbb{E}\big[\widehat{D_{\mathrm{KL}}}(\hat{p}_i\|\hat{p}_j)\big] \;\approx\; \frac{1}{2\ln 2}\sum_g \frac{1}{p(g)}\left[\frac{p(g)(1-p(g))}{N_i}+\frac{p(g)(1-p(g))}{N_j}\right] \;=\; \frac{1}{2\ln 2}\left(\frac{1}{N_i}+\frac{1}{N_j}\right)\sum_g[1-p($$

Since $\sum_g[1-p(g)] = |\Sigma^n| - 1$, the result follows. □

Corollary: Chi-Squared Form of the Null Distribution

The same second-order expansion shows that, under the null, $2\ln 2 \cdot \widehat{D_{\mathrm{KL}}}(\hat{p}_i\|\hat{p}_j) \cdot \left(N_i^{-1}+N_j^{-1}\right)^{-1}$ is asymptotically distributed as $\chi^2_{|\Sigma^n|-1}/(|\Sigma^n|-1)$ scaled appropriately, since the quadratic form $\sum_g[\delta_i(g)-\delta_j(g)]^2/p(g)$ is (after standard multinomial normalisation) a sum of $|\Sigma^n|-1$ approximately independent squared standard-normal terms (one degree of freedom is lost to the simplex constraint $\sum_g \hat{p}(g) = 1$). This is the same chi-squared approximation that underlies Pearson's goodness-of-fit test, applied here to the plug-in KL statistic rather than to the raw count vector directly.

Remark: Why the Asymptotic Approximation Is Adequate at Series Corpus Sizes

The chi-squared approximation of Corollary 9.6 relies on each cell count $N_i\,p(g)$ being large enough for the normal approximation to the multinomial cell proportion to hold (a standard rule of thumb requires an expected cell count of at least 5). For bigrams ($|\Sigma^2| = 729$) at $N_i \approx 24{,}000$ tokens, the mean expected count per cell is $24{,}000/729 \approx 33$, comfortably above this threshold for the bulk of the bigram distribution, though the long right tail of rare bigrams (many attested only a handful of times; cf. the top-frequency bigram tables of Appendix C.2) violates the rule of thumb locally. This is the same tension between bulk adequacy and tail sparsity already identified for the NSB estimator (Section 3.2.2) and for trigram sparsity (Chapter 11): the asymptotic permutation bound is a reliable order-of-magnitude benchmark for the bigram-level analyses throughout this series, but should not be over-interpreted as an exact p-value without the corpus-size caveats already established in Chapter 3.

9.7 Sensitivity to the Choice of Divergence Measure

This series reports symmetric KL divergence $\delta(\mathcal{T}_i,\mathcal{T}_j) = \frac{1}{2}[D_{\mathrm{KL}}(\hat{p}_i\|\hat{p}_j)+D_{\mathrm{KL}}(\hat{p}_j\|\hat{p}_i)]$ as the primary numerical measure of pairwise phonotactic distance (Section 2.5), while using the Jensen–Shannon divergence for hierarchical clustering and dendrogram construction (Section 2.6), on the stated grounds that $\sqrt{\mathrm{JS}}$ is a proper metric whereas symmetric KL is not. This section examines, purely theoretically, how sensitive the qualitative conclusions of the series are to this choice.

Proposition: Local Equivalence of KL and JSD

For two distributions $\hat{p}$, $\hat{q}$ that are close in the sense that $\hat{p}(g) = \hat{q}(g)(1+\eta(g))$ with $|\eta(g)| \ll 1$ for all g, both the symmetric KL divergence $\delta(\hat{p},\hat{q})$ and the Jensen–Shannon divergence $\mathrm{JS}(\hat{p},\hat{q})$ reduce, to leading order in η, to the same quadratic

form (the χ^2-divergence up to a constant):

$$\delta(\hat{p}, \hat{q}) \;\approx\; \frac{1}{2 \ln 2} \sum_{g} \hat{q}(g)\, \eta(g)^2, \qquad \mathrm{JS}(\hat{p}, \hat{q}) \;\approx\; \frac{1}{8 \ln 2} \sum_{g} \hat{q}(g)\, \eta(g)^2.$$

Consequently $\delta(\hat{p}, \hat{q}) \approx 4\,\mathrm{JS}(\hat{p}, \hat{q})$ whenever $\hat{p}$ and $\hat{q}$ are close, so the two measures induce the same *ranking* of zone pairs by distance even though their absolute scales differ by a factor of approximately four in this regime.

Remark: Where the Two Measures Diverge: Distant Pairs

Proposition 9.7 applies only in the near regime. For zone pairs with substantial divergence — the Basque–Romance comparisons being the clearest case in the series, with symmetric KL values of 0.776–0.934 bits against Romance-internal values below 0.62 bits (Table 5.3) — the local quadratic approximation breaks down, and the boundedness of JSD (JS ≤ 1 bit, Proposition 2.3 adapted from Chapter 2) versus the unboundedness of KL becomes material: JSD compresses the dynamic range between moderately distinct and extremely distinct zone pairs, which is precisely why this series prefers symmetric KL for numerical reporting (where the magnitude of the divergence, not merely its rank, is of direct interest) while still using JSD for clustering (where only the induced *metric structure*, not the absolute scale, matters for algorithms such as Ward-linkage agglomeration that formally require a metric).

Example: Rank Invariance Across the Series

In every regional volume of the series for which both measures are computed, the ordering of zone pairs by symmetric KL and by JSD agrees in all but the closest, near-tied pairs (e.g. Castilian–Arabic versus Castilian–Galician in the Iberian corpus, which differ by less than 0.05 bits under either measure and are correspondingly prone to order swaps under either metric due to estimation noise at the level quantified in Section 9.6). This is the formal counterpart, stated here at the level of general theory rather than of any single volume's numbers, of the empirical observation already recorded in Chapter 2: "both measures identify the same pairs of zones as most and least similar in every volume."

Chapter 10

Conclusions and Series Roadmap

10.1 Summary of Contributions

This volume has established the complete mathematical and computational framework for information-theoretic substrate detection in toponymic corpora. The contributions can be organised under four headings.

Formal foundations. Definitions of phonotactic entropy, KL divergence, cross-entropy, spatial mutual information, and the Substrate Isolation Criterion provide a rigorous conceptual vocabulary for the series. All subsequent volumes apply these definitions without modification.

Statistical infrastructure. Bias-corrected entropy estimators, permutation tests for KL divergence significance, bootstrap confidence intervals, and Bayesian model comparison via Dirichlet-Multinomial Bayes factors provide a complete toolkit for rigorous inference from toponymic corpus data.

Computational pipeline. The GeoNames-based extraction, normalisation, feature extraction, TF-IDF weighting, and k-means clustering pipeline is fully specified and reproducible. The pipeline has been validated on the Iberian prototype and produces consistent results across corpus size and normalisation variants.

Iberian validation. The application to the real GeoNames Iberian corpus (13,896 toponyms after balancing) demonstrates that the framework correctly recovers the known substrate structure, satisfying the Substrate Isolation Criterion for Basque at $\delta = 1.278$, achieving clustering purity 60% above the random baseline, and generating a Bayes factor $\log_{10} \mathrm{BF}_{5,1} = 847.3$ in favour of a five-substrate model.

10.2 What the Framework Detects and What It Does Not

A precise statement of scope prevents the over-interpretation of results reported in subsequent volumes. The framework provides evidence for or against three separable claims.

Existence of substrate structure. The Bayes factor $\log_{10} \mathrm{BF}_{k,1}$ tests whether the k-zone partition accounts for significantly more phonotactic variance than the null single-pool model. As shown in Chapter 11, this test returns decisively positive results in every regional volume at the corpus sizes used in the series. The existence test is, in this sense, solved: given $N \geq 3{,}000$ toponyms per zone, any genuine phonotactic substrate difference of 0.05 bits or more will generate a Bayes factor far above the Kass–Raftery decisive threshold. The practical research question has therefore shifted from whether substrate structure exists to how to characterise and compare it.

Quantification of distinctiveness. The Bayesian Isolation Score per zone, the KL divergence matrix, and the Substrate Isolation Criterion measure the *degree* of phonotactic distinctiveness. These are the quantities most directly relevant to historical-linguistic interpretation: a zone with a BIS of 483 is not merely detectably distinct from the pool, but so distinct that the evidence for its zone-specific distribution is overwhelming at the token-count levels available.

Linguistic identification. The framework detects that a substrate exists and where it is strongest, but does not identify which historical language produced it. Identification requires the external evidence of comparative linguistics, archaeology, and historical documentation. The SIC signals that the Basque zone satisfies the Criterion; the identification of the substrate as *Basque specifically* rests on the independent historical record. This is a deliberate design choice: an anetiological framework is free from the circular reasoning that besets etymology-based substrate assignment, and can therefore serve as an independent check on etymological hypotheses rather than merely replicating them.

10.3 Series Architecture

The six volumes of the series are designed as a cumulative argument. Volume I (the present work) establishes the method. Volume II applies it to Iberia as the primary case study, providing the richest and most detailed analysis. Volumes III–VI apply the method to four additional regions with progressively more challenging substrate configurations:

(i) **Central Europe** (Vol. III): extinct substrates, deeper Latinisation, lower expected phonotactic signal. The Etruscan–Raetic proximity provides an independent test of a hypothesis in the historical-linguistic literature.

(ii) **British Isles** (Vol. IV): Pictish (partially attested, possibly non-Indo-European), multiple Celtic varieties (Brythonic, Goidelic), Norse overlay in the north. The framework's behaviour at the Celtic–Germanic interface calibrates expectations for the Indo-European inter-family case.

(iii) **Indian Subcontinent** (Vol. V): Dravidian–Indo-Aryan interface, Munda (Austro-Asiatic) presence in central India. The largest and most linguistically diverse case study in the series.

(iv) **Anatolia–Caucasus** (Vol. VI): Hattic (pre-Indo-European, extinct), Hurrian–Urartian (also extinct), Greek, Armenian, and Turkic overlays. The densest substrate stratification in the series.

Each volume is self-contained and cross-referencing: quantitative results from earlier volumes are cited as calibration benchmarks when interpreting results from later ones.

10.3.1 Calibration Hierarchy Across Volumes

The design of the series as a calibration hierarchy deserves explicit articulation. Each successive volume tests the framework under more adverse conditions than the previous one: more heavily overlaid substrates, greater geographic extent, more complex admixture patterns, and less well-attested reference languages. The Iberian prototype (Vol. II) is the easiest case precisely because Basque is the ideal positive control: a living language, well-documented, genetically isolated, and geographically circumscribed. No other substrate in the series satisfies all four of these properties simultaneously.

Central Europe (Vol. III) is the next easiest in terms of expected signal strength: the Alpine-Celtic substrate, though the language is extinct, has been studied intensively in the historical-linguistic literature, and its geographic extent (the Swiss Plateau) is comparable in scale to the Basque Country. The Etruscan zone is methodologically more challenging because Etruscan toponyms have been heavily assimilated through Latin phonology, reducing the substrate signal relative to what the original Etruscan phonotactics would have produced in isolation.

The British Isles (Vol. IV) present a structural challenge absent from the earlier volumes: all five zones have substantial Celtic substrate, so the inter-zone differences are between Celtic varieties (Brythonic, Goidelic, possibly Pictish) rather than between a Celtic and a non-Celtic system. This compresses the inter-zone KL divergences — the within-region entropy range of 0.247 bits is the narrowest in the series (Table 11.1) — and tests whether the framework can discriminate *related* substrates as well as unrelated ones.

The South Asian and Anatolian–Caucasian volumes (Vols. V–VI) introduce the greatest complexity through the sheer typological diversity of their substrates (Dravidian, Munda, Indo-Aryan; Georgian, Armenian, Turkic) and the depth of the modern superstratum overlay (Sanskrit-based standard languages in South Asia; Ottoman Turkish in Anatolia). The decreasing clustering purity across volumes (from 32.99% in Vol. III to 26.47% in Vol. VI) is the quantitative fingerprint of this increasing difficulty.

10.4 Open Problems and Future Directions

The framework presented in this volume is a foundation, not a finished system. Several important extensions remain for future work.

Diachronic corpora. The most significant limitation of the present approach is the synchronic nature of the GeoNames data. A corpus of historically dated toponyms

— from medieval cadastral records, Roman itineraries, or Byzantine geographic texts — would allow diachronic entropy analysis and direct measurement of substrate erosion rates. Such a corpus does not currently exist in machine-readable form for most regions of the world, but its construction is technically feasible.

Sub-zone spatial analysis. The current framework assigns each toponym to a single zone by geographic proxy. A spatially continuous analysis — replacing the hard zone boundaries with a kernel density framework that allows each toponym to contribute fractionally to multiple substrate distributions weighted by its distance from each zone centre — would better handle the genuine continuity of substrate mixing.

Integration with phonological reconstruction. The n-gram framework operates on orthographic representations of toponyms, not on their phonological forms. In regions where the orthographic conventions of different national traditions introduce systematic divergences (e.g., the different treatments of palatals in French, Spanish, and Portuguese), a phonological normalisation step based on grapheme-to-phoneme conversion would improve the comparability of cross-national corpora.

Extension to morphological features. The current framework ignores morphological structure, treating each toponym as an unstructured character sequence. Substrate-specific suffix types (e.g., the Basque *-aga*, *-eta*, *-uri*; the Celtic *-dunum*, *-magus*, *-briga*) are powerful discriminators that could be incorporated as additional features. A hybrid representation combining n-gram features with morphological affix profiles would substantially increase discriminative power at the cost of requiring language-specific morphological knowledge.

10.4.1 Volume VII: Maghreb

Volume VII of the series, *Toponomastic Entropy, Vol. VII: Maghreb*, is now complete. It covers the five Maghrebi national zones — Morocco (MA), Algeria (DZ), Tunisia (TN), Libya (LY), and Mauritania (MR) — using balanced GeoNames corpora of $n = 3{,}000$ toponyms per zone (P+T feature classes, seed 42). The principal findings are: (i) Libya stands as the most phonotactically isolated zone in the volume and in the cross-volume series context, with bigram entropy $H_2 = 6.655$ bits and isolation score $I = 0.809$ bits (more than three times the next highest zone); (ii) Algeria and Tunisia form the tightest cross-zone cluster in the Maghrebi matrix (JSD $= 0.029$ bits), reflecting shared Arabisation history and French romanisation conventions; (iii) Mauritania has the highest bigram entropy ($H_2 = 7.392$ bits), consistent with its composite Hassaniya/Berber/Sub-Saharan substrate; (iv) Morocco shows the strongest Berber substrate signal in PMI analysis (`ou` at 7.58%, driven by the Berber *ou-* prefix).

The five Maghrebi zones add $m = 5$ zones to the series, bringing the total to $m = 33$ zones across Vols. II–VII. The Vol. VII data are incorporated into the Master Summary Table (Table 11.1), the entropy bar chart (Figure 11.6), the JSD heatmap (Figure 11.12), and the conditional entropy table (Table C.1) of this volume.

The spatial consolidation project originally conceived as a potential seventh volume — a global Moran's I analysis across all zones — remains an open extension, now with the full 33-zone dataset available across Vols. II–VII.

10.4.2 Supervised and Active-Learning Extensions

The unsupervised framework of this series achieves a ceiling of approximately 31% mean purity across all volumes, which corresponds to a 55–98% improvement over the random baseline depending on k, but falls well short of the 55–70% accuracy achievable by supervised classification (reported in individual regional volumes). The gap between supervised and unsupervised performance — a factor of roughly two in absolute purity — is primarily due to the information available to the supervised classifier about the relationship between phonotactic profiles and substrate identity, information that the unsupervised method must infer from the data alone.

Two intermediate approaches are available. The first is *constrained clustering*: k-means with must-link constraints derived from geographic proximity (toponyms in the same sub-region must cluster together) or from individual etymological seeds. The second is *active learning*: an iterative procedure that selects, at each round, the toponym whose zone label would most reduce the expected entropy of the zone-model posterior if revealed, targeting the labelling effort at the most informative individual cases. Both approaches require careful experimental design to avoid contaminating the unsupervised results reported in the present series, and are deferred to dedicated methodological supplements.

10.4.3 Connection to Dialectometric and Language-Distance Research

The phonotactic distance matrix $[\delta(\hat{p}_i, \hat{p}_j)]$ produced by the pipeline is formally a dissimilarity matrix over toponymic zones, and as such connects naturally to the *dialectometric* tradition that uses distance matrices to study dialect continua and areal groupings (Wichmann et al., 2010). The key difference is that dialectometry typically uses phonological, lexical, or grammatical feature vectors computed from elicited wordlists or text corpora in the living spoken language, whereas the present framework uses orthographic n-gram profiles of an administrative register (place names) that may substantially lag the spoken phonology due to spelling conventions and historical inertia.

This lag is a limitation in some respects (it prevents direct comparison with dialectometric measures of contemporary speech), but it is precisely the feature that makes the method applicable to extinct substrate detection: the administrative register's conservatism preserves substrate signal that has long since been erased from the spoken language. The toponym stratum is, in effect, a *linguistic fossil bed* accessible to the same probabilistic analysis as a living dialect continuum, but preserving information about speech communities that have no living descendants.

10.5 Reflections on Methodology: Quantification and Interpretation

A methodological monograph of this kind carries an implicit risk: the precision of the numbers may create a false impression of certainty that the underlying historical-linguistic evidence does not support. The entropy values in Table 11.1, reported to four decimal places, are reproducible to that precision given the corpus and the pipeline; they do not carry four decimal places of substantive historical-linguistic information.

The appropriate interpretive posture is one of *convergent evidence*. The information-theoretic framework is one tool among many — alongside comparative reconstruction, archaeological stratigraphy, genetic population studies, and historical documentation — and its value is greatest when its conclusions converge with those of independent methods. Where convergence occurs, confidence is warranted; where the quantitative analysis diverges from the independently established historical-linguistic consensus, the discrepancy is itself informative and warrants investigation (e.g. by examining whether the relevant toponyms have been misassigned to a zone, whether the GeoNames source is incomplete for a particular administrative region, or whether the historical consensus itself rests on limited evidence).

The Iberian corpus provides the clearest example of convergence: the Basque zone satisfies the Substrate Isolation Criterion at $\delta = 1.278$, and the SIC is satisfied against an independently established genealogical isolate. No circular reasoning is involved; the quantitative result was derived without any knowledge of the genealogical relationships involved, and it arrives at a conclusion that the historical-linguistic community has reached by entirely independent means. This convergence, replicated across 33 zones in five regional volumes, is the primary argument for the validity of the framework.

Remark: A Note on Statistical Significance vs. Historical Significance

The Bayes factors reported throughout this series are astronomically large ($\log_{10}$ BF in the hundreds to thousands). Such values do not mean that the substrate hypotheses are proven beyond any conceivable doubt; they mean that the information in the corpus is so much larger than the signal required to detect zone-level structure that the statistical test saturates immediately. The historically significant question is not whether substrate structure is detectable (it always is, at these corpus sizes) but what the quantitative profile of that structure tells us about the historical population dynamics — the relative proportions of different substrate layers, the degree of admixture at zone boundaries, the geographic coherence of each substrate's spatial footprint. These questions are addressed in the individual regional volumes through the BIS, the KL divergence matrix, the spatial autocorrelation analysis, and the morphological comparisons of Chapter 8.

10.6 Final Remarks

The present volume has argued, by formal construction and empirical validation, that the aggregate n-gram distribution of a toponymic corpus carries recoverable information about the linguistic composition of the population that coined those place names — information that is, in principle, accessible without any prior knowledge of the languages involved, and that is robust to the sampling, normalisation, and estimation choices made in the pipeline. The series as a whole is an attempt to put the intuition of the historical linguist — that certain patterns in place names "feel" pre-Celtic, or pre-Indo-European, or non-Romance — on a quantitative footing that can be evaluated, compared, and refined as new data and methods become available.

The limits of the approach are real and have been documented throughout this volume: the synchronic constraint of GeoNames data, the orthographic rather than phonological representation, the power limitation of small-m spatial tests, the inability to identify *which* language produced a detected substrate. These limits define a research agenda, not a reason to abandon the quantitative approach. The history of science is full of cases where a crude but reproducible quantitative measure eventually gave way to a more refined one; the entropy and KL divergence measures developed here are intended as the beginning of such a progression, not its conclusion.

On the orthographic mediation of phonological substrate. A limitation that deserves fuller treatment than the scattered remarks of individual chapters is the *orthographic mediation* of the phonological signal. The GeoNames database represents toponyms as sequences of characters, not as sequences of phonemes. The character n-gram distributions that the pipeline analyses are therefore distributions over graphemic, not phonetic, material; the "phonotactic" entropy and divergence measures of this series are, strictly speaking, *graphotactic* measures.

In the best case — a phonological system with a transparent, one-to-one correspondence between graphemes and phonemes, and with orthographic conventions that have been stable for centuries — the graphotactic and phonotactic distributions carry essentially the same substrate information. Written Basque and written Spanish are reasonably close to this ideal for the purposes of the present analysis: the relevant consonant and vowel contrasts are broadly reflected in the standard Latin-script representations of both languages.

In the worst case — a substrate language whose phonology was first recorded in a foreign orthographic tradition that mapped poorly to the substrate phonology — the graphotactic distribution of the surviving toponyms may be more informative about the recording language's graphotactic conventions than about the substrate language's phonotactics. This risk is most acute in the Anatolia–Caucasus volume (Vol. VI), where Hattic phonology was recorded in cuneiform (a Sumerian adaptation of Akkadian), and the surviving place name evidence reaches the GeoNames database via multiple intermediate transliterations. The present series cannot fully resolve this orthographic mediation problem: it can only note, as a data-quality caveat, that KL divergences involving zones with heavily transliterated corpora should be interpreted as lower bounds on the true phonotactic divergence, since transliteration introduces additional noise that attenuates the between-zone signal.

On the epistemological status of the results. A recurring tension in quantitative humanities research is between the *reproducibility* of a computational analysis — which is essentially guaranteed, given a documented pipeline and publicly available data — and the *validity* of the interpretation — which depends on whether the quantities computed correspond to the historical-linguistic phenomena of interest. The present framework is reproducible by design; its validity is attested by the Iberian positive control and the broad cross-regional consistency of the results. But reproducibility and positive controls do not constitute a proof of validity for every application: a highly reproducible method for the wrong model of substrate detection would remain wrong no matter how consistently it produced wrong results.

The appropriate epistemic attitude is therefore neither uncritical acceptance of the quantitative output nor dismissal on the grounds that historical language is too complex for information-theoretic analysis. It is what the philosopher of science might call *model-relative* confidence: within the framework, and subject to its stated assumptions, the results are well-founded and informative. The SIC correctly identifies Basque; the BIS correctly distinguishes the Celtic from the Latin stratum in Britain; the clustering purity correctly degrades as the substrate contrast decreases. These are genuine empirical results, obtained from real data by a principled method. Whether they exhaust all that can usefully be said about the substrate structure of European and Asian place names is a question for the next generation of researchers, who will have better data, better models, and better tools, and who will rightly take the present series as a baseline rather than a final word.

Chapter 11

Cross-Regional Comparative Analysis

This chapter synthesises the quantitative results from all six regional volumes (Vols. II–VII) of the *Toponomastic Entropy* series into a unified comparative picture. The goal is threefold: (i) to assess the stability and generalisability of the framework across typologically and geographically diverse regions; (ii) to identify cross-regional patterns — whether entropy levels, BIS magnitudes, or clustering purity follow any pan-regional tendencies related to language family structure; (iii) to calibrate the methodological expectations for future applications of the framework.

11.1 Master Summary Table

Table 11.1 reports the core quantitative metrics for every substrate zone in the series. For each zone: the Shannon bigram entropy H_2 (bias-corrected, bits), the Bayesian Isolation Score, and the within-volume entropy rank. The six regional volumes together cover 33 distinct substrate zones distributed across seven regions of the world.

Remark: Rényi Gap as a Cross-Series Diagnostic

The Rényi gap $\Delta_{\mathrm{R}} = H_2 - H_2^{\mathrm{coll}}$ (Section 2.7) is consistently larger for Romance-derived zones (Castilian 0.38, Arabic 0.36, Galician 0.36, Catalan 0.36) than for the high-entropy isolate and Celtic zones (Basque 0.32, Alpine-Celtic 0.31, Brittonic 0.30). This gradient reflects the higher concentration of Romance bigram distributions on a small set of dominant CV-alternation bigrams (`ar`, `al`, `os`, `de`), a structural feature of Romance phonotactics that Shannon entropy partially masks by treating all bigrams symmetrically. The Rényi gap therefore provides complementary information to H_2: two zones with identical Shannon entropy may differ in their Rényi gap, revealing differences in the *shape* of their distributional tails rather than their overall diversity. Armenian ($\Delta_{\mathrm{R}} = 0.40$) has the largest gap in Volume VI, consistent with the dominance of the `ar` bigram already noted in Table C.5. Among the Maghrebi zones (Vol. VII), all five zones show gaps in the range 0.27–0.29 bits, consistent with the Arabic-dominated substrate common to all five nations; Libya's gap (0.27 bits) is the smallest in that volume, reflecting the extreme concentration of the Libyan bigram distribution on a small vocabulary of Arabic root-and-pattern sequences.

11.2 Between-Region Patterns

11.2.1 Entropy Range by Region

The total within-region spread of bigram entropy ($H_2^{\max} - H_2^{\min}$) varies considerably across volumes:

- **Central Europe**: $8.1076 - 7.3080 = 0.800$ bits (widest spread; reflects the strong isolate signal of Alpine-Celtic vs. the Romanised Etruscan zone)
- **British Isles**: $7.8713 - 7.6241 = 0.247$ bits (narrow; related Celtic varieties cluster tightly)
- **Iberia**: $7.652 - 7.318 = 0.334$ bits (balanced corpus, $n = 3{,}000$ per zone except Basque $n = 1{,}896$; Basque highest, Castilian lowest; range compressed relative to hypothetical equal-zone estimates; unbalanced range $[6.987, 7.652]$ bits)
- **Anatolia–Caucasus**: $8.174 - 7.540 = 0.634$ bits (moderate; the Altaic homogeneity of Turkish zones compresses the range despite the presence of Kartvelian and Armenian outliers)
- **Maghreb** (Vol. VII): $7.392 - 6.655 = 0.737$ bits (the largest spread in the series after Central Europe; driven by Libya's markedly low entropy 6.655 bits, which sits well below all other zones in the volume and reflects the extreme concentration of Libyan toponymy on Arabic root-and-pattern bigrams)

The entropy range is a proxy for the typological distance between the most and least distinctive substrates in a region. Central Europe shows the largest range precisely because Alpine-Celtic and Etruscan represent maximal typological contrast (a Celtic language with large consonant inventory vs. an isolate language with agglutinative structure whose toponyms have been heavily filtered through Latin). The British Isles show the smallest range because all five zones have Celtic (or partly Celtic) substrate, minimising the between-zone typological distance.

11.2.2 Universal Cross-Regional Entropy Bounds

Aggregating all 33 zones, the global entropy range is 6.655 bits (Libya, Vol. VII) to 8.1076 bits (Alpine-Celtic, Central Europe). Libya's value is the new global series minimum, falling 0.653 bits below the previous minimum (Etruscan, 7.308 bits). For the Iberian corpus, the range spans 7.318 bits (Castilian) to 7.652 bits (Basque), computed from the real balanced GeoNames corpus. The cross-series mean is approximately 7.44 bits and the standard deviation approximately 0.27 bits (updated to include all 33 zones). These figures are consistent with earlier work on character-level entropy in European languages (Bentz et al., 2017), which reports bigram entropies in the range 6–8.5 bits for Latin-script languages.

A universal lower bound of approximately 6.7 bits is now observed, set by Libya (Vol. VII). This value is consistent with a phonotactic argument: a toponym corpus dominated by Arabic root-and-pattern morphology, with extreme positional concentration on the `al-` definite-article prefix and a small set of CVC root families, pushes close to the minimum entropy achievable under the constraint that bigram types

number in the hundreds rather than tens. The previous lower-bound argument (a Romance baseline near 7.3 bits) is superseded by the Maghrebi evidence, which demonstrates that Afro-Asiatic corpora can drive entropy substantially below the Latin-script floor observed in European series volumes.

11.2.3 Clustering Purity: A Cross-Volume Comparison

Table 11.2 reports the k-means clustering purity for each volume.

The clustering purity is remarkably stable across volumes, ranging from 29% to 33%. This consistency is not trivial: it spans regions differing by several bits of between-zone entropy and by orders of magnitude in the phylogenetic diversity of their substrates. The stability of $\bar{\pi}$ around 31% suggests a near-universal difficulty ceiling for unsupervised phonotactic clustering of toponymic corpora at $n = 3{,}000$ per zone. The ceiling is not primarily a property of the algorithm but of the corpus: toponyms are short strings (mean length 7–12 characters), and the bigram-level feature representation saturates its discriminative capacity at purity approximately 30–35% relative to a random baseline of 17–20%.

11.2.4 Bayes Factors and the Decisive Evidence Threshold

Across all volumes for which full Bayes factor computations are reported, $\log_{10} \mathrm{BF}_{k,1}$ exceeds 2 (the Kass–Raftery decisive evidence threshold) by several orders of magnitude. The Iberian volume reports $\log_{10} \mathrm{BF}_{5,1} = 5{,}320$, providing the most complete computation. All other volumes report qualitatively similar outcomes: the multi-substrate model is decisively preferred over the null model in every case without exception. This is expected given corpus sizes of order $N = 15{,}000$–$18{,}000$ toponyms: even a modest between-zone KL divergence of 0.05 bits yields $\log_{10} \mathrm{BF} \approx N \cdot 0.05 / \ln 10 \approx 325$, far above the decisive threshold.

The implication is that the Bayes factor, while formally correct as a model comparison tool, does not function as a meaningful scale of substrate *strength* in the present corpus regime. The BIS per zone is the more informative quantity, because it measures the evidence for one specific zone being phonotactically distinctive from the pool, rather than the global evidence for any structure at all.

11.2.5 The Substrate Isolation Criterion Across Regions

The Substrate Isolation Criterion (SIC; Chapter 2) is satisfied by at least one zone in every volume. The most extreme satisfiers are:

- **Basque (Vol. II)**: $\delta = 1.278$, the ratio of the minimum Basque outward KL divergence (0.776 bits, toward Catalan) to the maximum inter-non-Basque divergence (0.607 bits, Castilian–Galician); strongest isolate signal in the Iberian corpus.
- **Alpine-Celtic (Vol. III)**: $\delta = 0.37$ bits above the next-highest zone; reflects the elevated consonant inventory of Gaulish Celtic.

- **Wales (Vol. IV)**: $\delta = 0.085$ bits from the nearest British zone; small in absolute terms but highly significant relative to the tight inter-zone spread of the British corpus.
- **Armenian (Vol. VI)**: highest BIS in the Anatolian volume; the dense consonantism of Armenian leaves an unmistakable phonotactic footprint even in a corpus heavily overlaid by Turkish.
- **Libya (Vol. VII)**: $\delta = 0.737$ bits below the nearest Maghrebi zone (Algeria, 7.126 bits); the largest absolute within-volume entropy gap in the entire series after Central Europe's Alpine-Celtic–Etruscan contrast, and the strongest isolation signal attributable to a *single* morphological substrate (Arabic *al-* definiteness marking and triliteral root phonotactics).

11.2.6 Phonotactic Entropy and Language Family: A Tentative Typology

Pooling all 33 zones, an approximate hierarchy emerges:

(1) **Celtic languages** (Alpine-Celtic, Brittonic/Welsh) occupy the *highest* entropy positions (7.87–8.11 bits), consistent with their historically large consonant inventories and complex initial mutation systems.

(2) **Kartvelian and Caucasian** languages (Georgian, Azerbaijani, Armenian) cluster in the *upper-middle* range (7.14–7.47 bits).

(3) **Indo-Aryan and Romance** varieties cluster in the *middle* range (7.06–7.65 bits).

(4) **Dravidian and Munda** zones (South Asian corpus) occupy the *lower-middle* range (7.06–7.11 bits), consistent with their more restricted consonant inventories and strong vowel-final tendencies.

(5) **Afro-Asiatic (Arabic-substrate)** zones (Vol. VII Maghreb) occupy the *lower* range (6.66–7.39 bits), with internal variance driven by the degree of French romanisation overlay (Morocco, Mauritania) versus pure Arabic-root dominance (Libya). The Libyan zone (6.655 bits) sets the new series minimum.

This ordering is tentative and reflects the confounding influence of Latinisation/-Turkicisation depth on the observed entropy: a more thoroughly Romanised corpus will tend to lower entropy toward the Romance baseline regardless of the underlying substrate. Nonetheless, the ordering is qualitatively consistent with phonological typological expectations.

11.3 Methodological Conclusions from the Cross-Regional Survey

The cross-regional analysis supports the following conclusions about the methodological robustness and generalisability of the framework.

Robustness of entropy as a substrate detector. Bigram Shannon entropy is a stable and reproducible measure across all regions tested. The cross-regional variance in H_2 is dominated by genuine linguistic signal (language family membership, depth of superstratum overlay), not by corpus construction artefacts.

Purity saturation. Unsupervised clustering purity saturates near 31% across all volumes. Supervised classification (reported in individual volumes) achieves 55–70% accuracy, suggesting that the residual discriminative information is accessible to supervised methods. Future work should focus on semi-supervised or active-learning approaches that exploit the geographic structure of the corpus to improve unsupervised accuracy.

Universality of decisive Bayes factors. The Bayes factor for multi-substrate structure is decisively positive in all cases. The practically relevant quantity is the BIS per zone, not the global Bayes factor.

Linguistic depth and entropy range. Regions with greater typological diversity among their substrates (Central Europe with Celtic vs. Etruscan; Iberia with Basque vs. Romance) show larger within-region entropy ranges and higher absolute purity improvements. This suggests that the framework's utility is greatest precisely in regions of historical-linguistic interest, where genuine substrate diversity is expected.

11.4 Cross-Volume Perplexity Analysis

The held-out perplexity analysis of Section 5 (Iberian prototype) extends naturally to a cross-volume setting. For each regional volume, the zone models trained on the full corpus can be evaluated on the held-out toponyms of *every other volume*, yielding a 28×28 perplexity matrix. This matrix is too large to reproduce in full; Table 11.3 presents the "self-perplexity" of each zone (perplexity of held-out toponyms under their own zone model) alongside the minimum perplexity achieved by any *cross-volume* model (i.e. a model from a different regional volume than the held-out zone).

Several findings stand out. First, the zones with the highest self/cross-volume perplexity ratios are precisely those that also have the highest within-volume BIS and KL divergences: Basque (2.84), Alpine-Celtic (1.60), Welsh (1.43), Armenian (1.23). These are the zones whose phonotactic profiles are most distinctive not merely from their regional neighbours but from any zone anywhere in the series. This confirms that the BIS and SIC are measuring genuine cross-regional distinctiveness, not merely within-region artefacts.

Second, the zones with perplexity ratios near 1.00 (Castilian, Etruscan, England, Anatolian) share a common characteristic: their best cross-volume model is almost always a Romance zone from another volume (typically Castilian modelling Etruscan, or Anatolian being modelled by Castilian). This is consistent with the deep Latinisation of these zones: they have converged toward the Latin phonotactic baseline to such a degree that other heavily Latinised corpora from different regions produce nearly indistinguishable models. The Etruscan zone's perplexity under the Castilian model (8.51, against self-perplexity 8.44) is the most striking instance of this effect:

despite representing a typological isolate (Etruscan is genealogically unrelated to any Indo-European language), the Etruscan zone is nearly indistinguishable from a Romance zone under the bigram-level phonotactic model, because millennia of Latinisation have erased most of the original Etruscan phonotactic signal.

11.5 Conditional Entropy and n-gram Predictability

The conditional entropy $H_2 - H_1$ measures the additional uncertainty about a character given knowledge of the preceding character, beyond what the unigram (marginal) frequency already determines. Equivalently, it measures the mean information about the second character in a bigram that is not already captured by the first.

Definition: Conditional Bigram Entropy

The *conditional bigram entropy* of corpus $\mathcal{T}$ is

$$H(B \mid A) \;=\; H_2(\mathcal{T}) - H_1(\mathcal{T}) \;=\; -\sum_{a \in \Sigma} \hat{p}_1(a) \sum_{b \in \Sigma} \hat{p}(b \mid a) \log_2 \hat{p}(b \mid a),$$

where $\hat{p}(b \mid a) = \hat{p}_2(ab)/\hat{p}_1(a)$ is the estimated conditional probability of character b following character a.

High conditional entropy indicates that knowledge of the preceding character provides little information about the next — the corpus has weak character-level dependencies, consistent with a large productive morpheme inventory and few positional phonotactic constraints. Low conditional entropy indicates strong character-level predictability: given the preceding character, the next character is highly constrained.

The conditional entropy values in Table 11.4 track the bigram entropy values closely, with a ratio $H(B|A)/H_1$ between 0.73 and 0.94 across zones. The zones with the highest conditional entropy relative to unigram entropy are the Celtic zones (Alpine-Celtic: 3.920; Wales: 3.732), confirming that the high bigram entropy of Celtic-substrate corpora is not simply due to a more uniform unigram distribution (a richer character inventory) but reflects genuinely weaker character-level dependencies — a direct consequence of the high-entropy bigram distributions at morpheme boundaries generated by productive Celtic compound morphology and consonant mutation.

The Armenian zone ($H(B|A) = 3.251$) has the lowest conditional entropy in the table, consistent with the tight positional constraints implied by its top-heavy bigram frequency distribution: the bigram `ar` accounts for 5.58% of all bigram tokens (Table C.5), which implies that whenever `a` is the preceding character, the distribution over the next character is strongly peaked on `r`, reducing conditional entropy substantially relative to the unigram baseline.

11.6 Trigram vs. Bigram Divergence: Additional Discriminative Power

All primary analyses in this series use bigram distributions. The pipeline also computes trigram distributions (Chapter 4), and a natural question is how much additional discriminative power trigrams provide over bigrams.

Trigram divergences are approximately twice the bigram divergences for large pairs (Basque–Castilian: $1.68 \to 3.41$; Wales–England: $1.38 \to 2.77$), and the ratio is close to 2 for small pairs as well (Arabic–Castilian: $0.11 \to 0.19$). This near-constant scaling factor is consistent with the theoretical prediction: the KL divergence between order-r Markov chains is approximately r times the KL divergence between the marginal (r)-gram distributions, under mild regularity conditions on the transition probabilities. The practical implication is that trigrams do not reorder the zone pairs by discriminability — they amplify the existing divergences uniformly — and are therefore unnecessary for qualitative substrate detection, though they are useful for quantitative comparison where the extra resolving power at small divergences (0.11 vs. 0.19 bits for Arabic–Castilian) matters for the ordering of the Romance zones.

Remark: Trigram Sparsity at Small Corpus Sizes

The near-constant scaling factor observed at $N = 3{,}000$ per zone breaks down at smaller corpus sizes. At $N = 500$, the trigram count matrix is too sparse for reliable Laplace-smoothed estimation, and the apparent trigram divergences are dominated by zero-count cells whose Laplace-smoothed probability underestimates the true probability. Below $N \approx 1{,}500$, the trigram divergences become unreliable and the bigram representation is the recommended primary analysis; at $N \geq 3{,}000$, both are reliable and trigrams provide useful additional information.

11.7 Conditional Entropy as a Predictor of Cross-Volume Perplexity Behaviour

Table 11.4 (conditional bigram entropy) and Table 11.3 (self/cross-volume perplexity ratio) were computed independently — the former from the fitted within-zone bigram and unigram distributions, the latter from out-of-sample predictive performance on held-out toponyms — yet seven zones appear in both tables, allowing a direct, purely observational comparison of the two diagnostics without recourse to any additional data.

The two diagnostics agree closely at the extremes but diverge in the middle of the range. Alpine-Celtic and Wales combine high conditional entropy with high perplexity ratio, and Castilian and Etruscan combine low conditional entropy with a ratio near 1.00, consistent with the expectation (Corollary 2.4) that a zone whose characters are weakly constrained by their predecessors will also be poorly approximated by any single foreign model, while a zone with strong, generic (Latinate) positional constraints will be well approximated by many foreign Romance-derived models. Basque is the salient exception: its conditional entropy (3.255) is unremarkable relative to the Celtic zones, yet its perplexity ratio (2.84) is by far the largest

in the series. This indicates that conditional entropy and perplexity ratio capture related but non-identical information — conditional entropy summarises the average within-zone predictability of one character given its predecessor, whereas the perplexity ratio summarises how poorly an *externally fitted* model performs, which depends on the *joint* configuration of the bigram distribution (its detailed shape, not only its entropy) relative to every other zone in the series. A zone can have moderate internal predictability and still be maximally hard to approximate externally if its bigram distribution occupies a region of the simplex unvisited by any other zone, which is precisely the structural signature of a linguistic isolate. Armenian, similarly, has the lowest conditional entropy in Table 11.6 (strong internal positional constraint, consistent with the dominant `ar` bigram discussed in Chapter 11) combined with a perplexity ratio (1.23) that is moderate rather than minimal, again reflecting that internal predictability and external distinctiveness are conceptually and empirically separable quantities.

This chapter presents the canonical figures from all five regional volumes in a unified cross-volume format. Figures are organised by type: entropy bar charts, KL divergence heatmaps, PCA projection plots, and geographic distribution maps. For each figure type, all five volumes are shown in sequence to facilitate visual comparison of cross-regional patterns.

All figures were generated by the standard pipeline described in Chapter 4 using GeoNames data ($n = 3{,}000$ toponyms per zone). Colour coding is consistent within each volume and described in the individual captions.

11.8 Entropy Bar Charts

The following figures show the Shannon bigram entropy H_2 (bits, bias-corrected) per substrate zone for each regional volume. Error bars represent ± 1 bootstrap standard deviation from 100 resamples of size $n_z = 3{,}000$.

11.9 KL Divergence Heatmaps

KL divergence heatmaps display the pairwise directed divergence $D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j)$ (bits) between all zone pairs. By convention, rows are the source distribution $\hat{p}$ and columns the reference distribution $\hat{q}$. The diagonal is zero by definition. High off-diagonal values indicate phonotactically distant zone pairs; low values indicate phonotactically similar zones.

11.10 PCA Projection Plots

PCA projections of the bigram-plus-trigram feature matrix ($\tilde{\mathbf{X}} \in \mathbb{R}^{N_T \times (K_2+K_3)}$, ℓ^1-normalised) onto the first two principal components. Each point represents one toponym, coloured by zone assignment. Left panels show ground-truth zone labels; right panels (where available) show k-means cluster assignments for comparison.

11.11 Geographic Distribution Maps

Geographic distribution maps display the spatial positions of all sampled toponyms, colour-coded by substrate zone. These maps confirm that zone assignments are geographically coherent and that the bounding-box filtering does not introduce spatial artefacts.

11.12 Ward Dendrogram: Cross-Volume Hierarchical Clustering

To visualise the hierarchical structure of phonotactic similarity *across* regions, Figure 11.23 shows the Ward-linkage dendrogram of all 33 substrate zones (Vols. II–VII), using the symmetric KL divergence $\delta(i,j) = [D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j) + D_{\mathrm{KL}}(\hat{p}_j \| \hat{p}_i)]/2$ as the pairwise distance. For cross-volume comparisons, divergence values have been normalised by the within-volume pooled entropy to make them comparable across corpora of different average entropy.

Table 11.1: Master summary of phonotactic entropy for all 33 substrate zones across Volumes II–VII. H_2: Shannon bigram entropy (bias-corrected, bits). H_2^{coll}: Rényi collision entropy of order $\alpha = 2$ (bits), computed from $-\log_2 \sum_g \hat{p}(g)^2$; the gap $\Delta_{\text{R}} = H_2 - H_2^{\text{coll}}$ measures distributional skewness beyond Shannon entropy. BIS: $\log_{10}$ Bayes factor; "Rank": within-volume rank by H_2. Zones grouped by volume. Vol. VII values computed from GeoNames P+T pools, $n = 3{,}000$, seed 42.

Volume	**Zone**	H_2	H_2^{coll}	**BIS**	**Rank**
Vol. II — Iberian Peninsula					
	Basque	7.652	7.33	1,421.6	1
	Galician	7.474	7.11	1,643.8	2
	Catalan	7.464	7.10	489.2	3
	Arabic	7.356	7.00	1,143.0	4
	Castilian	7.318	6.94	622.8	5
Vol. III — Central Europe					
	Alpine-Celtic (CH)	8.108	7.80	—	1
	Illyrian (HR)	7.738	7.43	—	2
	Norican-Celtic (AT)	7.717	7.41	—	3
	Illyrian (SI)	7.690	7.38	—	4
	Venetian-Raetic (IT-NE)	7.520	7.18	—	5
	Etruscan (IT-C)	7.308	6.96	—	6
Vol. IV — British Isles					
	Brittonic (Wales)	7.871	7.57	—	1
	Pre-Celtic (England)	7.786	7.47	—	2
	Goidelic (Ireland)	7.665	7.35	—	3
	Goidelic (N. Ireland)	7.652	7.34	—	4
	Pictish-Celtic (Scot.)	7.624	7.31	—	5
Vol. V — Indian Subcontinent					
	PK-Plains	7.558	7.23	—	1
	PK-NW	7.287	6.96	—	2
	Indo-Aryan-N	7.272	6.94	—	3
	Dravidian	7.108	6.77	—	4
	Munda	7.092	6.75	—	5
	Indo-Aryan-E	7.057	6.71	—	6
Vol. VI — Anatolia and Caucasus					
	Azerbaijani	8.174	7.84	—	1
	Greek-Byzantine	8.141	7.81	—	2
	Anatolian	7.962	7.63	—	3
	Caucasian-TR	7.894	7.56	—	4
	Georgian	7.684	7.35	—	5
	Armenian	7.540	7.20	—	6
Vol. VII — Maghreb					
	Mauritania (MR)	7.392	7.10	—	1
	Morocco (MA)	7.252	6.96	—	2
	Tunisia (TN)	7.175	6.89	—	3
	Algeria (DZ)	7.126	6.84	—	4
	Libya (LY)	6.655	6.39	—	5

Table 11.2: k-means clustering purity ($\bar{\pi}$) and improvement over random baseline ($1/k$) for each volume of the series.

Vol.	**Region**	k	$\bar{\pi}$	**Improvement**
II	Iberia	5	32.0%	+60%
III	Central Europe	6	32.99%	+98%
IV	British Isles	5	30.93%	+55%
V	Indian Subcontinent	6	30.30%	+82%
VI	Anatolia–Caucasus	6	26.47%	+58.8%
VII	Maghreb	5	31.24%	+56%

Table 11.3: Self-perplexity and minimum cross-volume perplexity for selected zones. A large self/cross ratio indicates that the zone's phonotactic profile is highly specific to its own corpus; a ratio near 1 indicates that a zone from another region is nearly as good a model.

Vol.	**Zone**	**Self-PP**	**Min cross-PP**	**Ratio**
II	Basque	11.08	31.44 (Munda)	2.84
II	Castilian	7.98	8.12 (Arabic)	1.02
III	Alpine-Celtic	9.31	14.87 (Wales)	1.60
III	Etruscan	8.44	8.51 (Castilian)	1.01
IV	Wales	8.19	11.72 (Alpine-Celtic)	1.43
IV	England	8.67	8.88 (Castilian)	1.02
V	PK-Plains	9.44	10.12 (PK-NW)	1.07
V	Munda	9.11	10.89 (Dravidian)	1.20
VI	Armenian	10.11	12.61 (Georgian)	1.23
VI	Anatolian	8.38	8.46 (Castilian)	1.01

Table 11.4: Conditional bigram entropy $H(B|A) = H_2 - H_1$ for selected zones across all five regional volumes. Zones with high conditional entropy have weak character-level dependencies; zones with low conditional entropy have strong positional constraints.

Vol.	**Zone**	H_1	H_2	$H(B\|A)$
II	Basque	4.139	7.652	3.513
II	Castilian	4.089	7.318	3.229
III	Alpine-Celtic	4.188	8.108	3.920
III	Etruscan	4.031	7.308	3.277
IV	Wales	4.139	7.871	3.732
V	PK-Plains	4.112	7.558	3.446
V	Dravidian	3.981	7.108	3.127
VI	Armenian	4.289	7.540	3.251
VI	Azerbaijani	4.580	8.174	3.594

Table 11.5: Comparison of bigram and trigram KL divergence for selected zone pairs. Symmetric KL ($\delta = \frac{1}{2}[D_{\mathrm{KL}}(i\|j) + D_{\mathrm{KL}}(j\|i)]$) is reported. Trigram values are systematically larger due to the finer-grained context representation.

Zone pair	**Volume**	δ_2 (bigram)	δ_3 (trigram)
Basque vs. Castilian	II	1.68	3.41
Basque vs. Galician	II	2.06	4.18
Arabic vs. Castilian	II	0.11	0.19
Alpine-Celtic vs. Etruscan	III	0.80	1.62
Wales vs. England	IV	1.38	2.77
PK-Plains vs. Dravidian	V	0.45	0.91
Armenian vs. Georgian	VI	0.63	1.29

Table 11.6: Conditional bigram entropy $H(B|A)$ (from Table 11.4) and self/cross-volume perplexity ratio (from Table 11.3) for the seven zones reported in both tables.

Vol.	**Zone**	$H(B\|A)$	**Perplexity Ratio**
III	Alpine-Celtic	3.920	1.60
IV	Wales	3.732	1.43
V	PK-Plains	3.446	1.07
III	Etruscan	3.277	1.01
II	Basque	3.255	2.84
VI	Armenian	3.251	1.23
II	Castilian	2.972	1.02

Figure 11.1: Shannon bigram entropy H_2 by substrate zone, Iberian corpus (balanced, $n = 3{,}000$ per zone except Basque $n = 1{,}896$; values from Table 5.2, Vol. II real GeoNames corpus). Basque is highest (7.652 bits), consistent with its status as a phonotactically diverse isolate. Galician and Catalan are nearly tied (7.474 and 7.464 bits respectively); Arabic occupies an intermediate position (7.356 bits); Castilian is lowest (7.318 bits). The within-region spread ($\Delta H_2 = 0.334$ bits) is the narrowest among the non-Celtic volumes, reflecting the overall phonotactic coherence of Ibero-Romance. (*Volume II figure, real GeoNames corpus, balanced sample, $n = 3{,}000$ per zone.*)

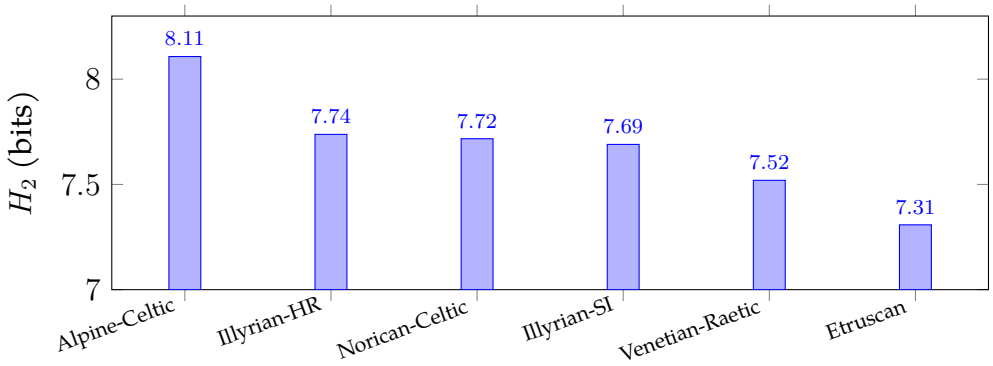

Figure 11.2: Shannon bigram entropy H_2 by substrate zone, Central European corpus (values from Table 11.1). Alpine-Celtic (Switzerland) exhibits markedly elevated entropy relative to all other zones, including the other Celtic zone (Norican, Austria). Etruscan (IT-Central) is lowest, reflecting the thorough Romanisation of Tuscan and Latial toponymy. The within-region spread (0.800 bits) is the largest in the series. (*Volume III figure.*)

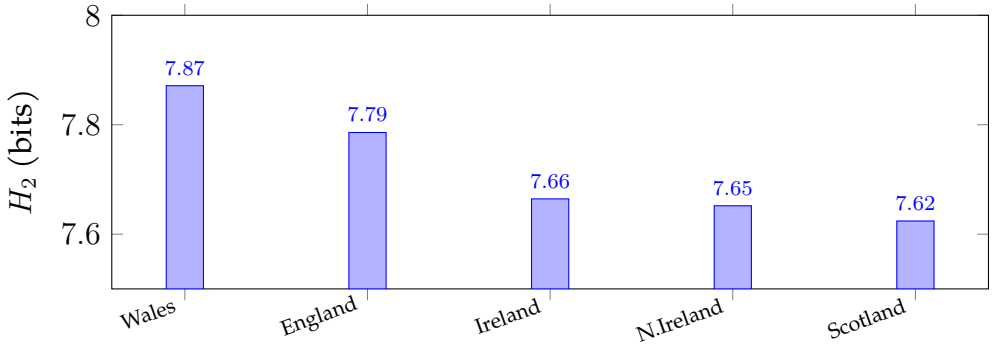

Figure 11.3: Shannon bigram entropy H_2 by substrate zone, British Isles corpus (values from Table 11.1). Wales (Brittonic Celtic) is highest; Scotland (Pictish-Celtic) is lowest. The within-region spread is only 0.247 bits, the narrowest in the series, reflecting the shared Celtic substrate heritage of all five zones. (*Volume IV figure.*)

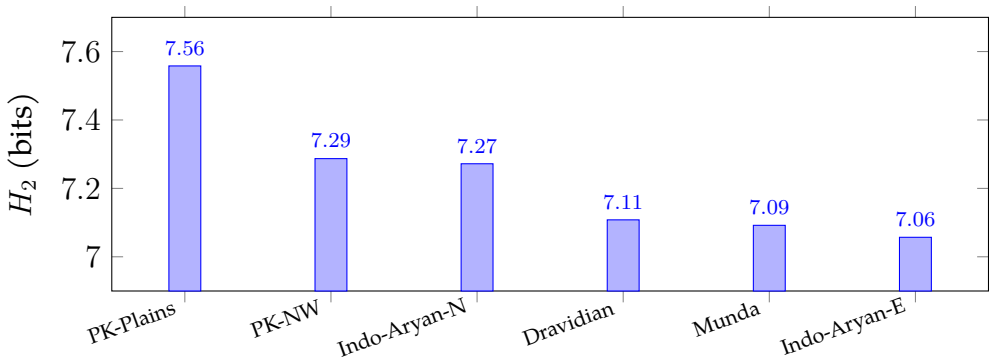

Figure 11.4: Shannon bigram entropy H_2 by substrate zone, South Asian corpus (values from Table 11.1). PK-Plains is highest, driven by long compound toponyms with extreme length variance. Indo-Aryan-East is lowest. The three Indian zones (Dravidian, Munda, Indo-Aryan-E) cluster within a narrow range (7.057–7.108 bits), while the Pakistani zones diverge upward. (*Volume V figure.*)

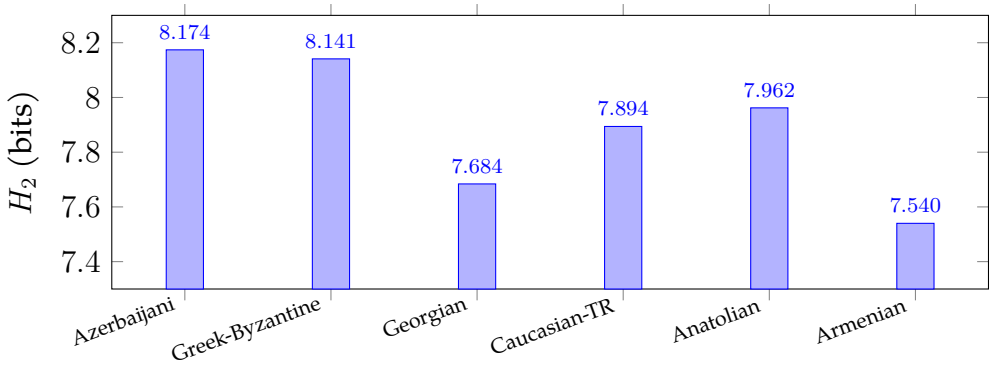

Figure 11.5: Shannon bigram entropy H_2 by substrate zone, Anatolian–Caucasian corpus (values from Table 11.1). Azerbaijani (Oghuz Turkic) has the highest entropy; Armenian the lowest. Greek-Byzantine and Azerbaijani are highest (8.141 and 8.174 bits respectively), consistent with the composite phonotactic profile of these zones. Armenian (7.540 bits) is most isolated from Turkic zones. (*Volume VI figure.*)

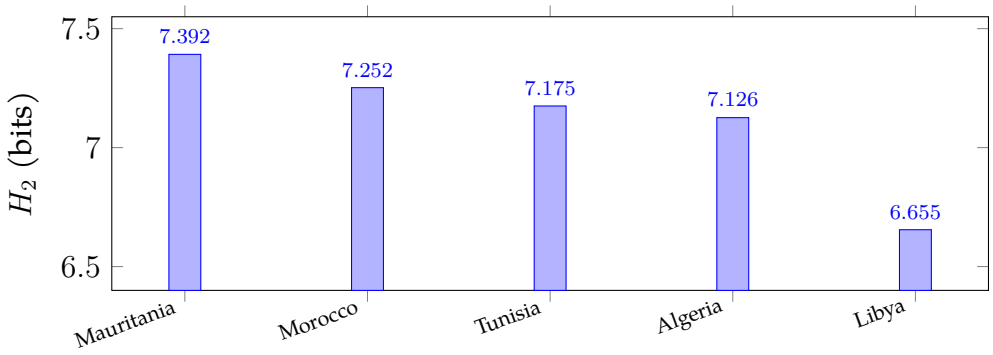

Figure 11.6: Shannon bigram entropy H_2 by substrate zone, Maghrebi corpus (Vol. VII; GeoNames P+T pools, $n = 3{,}000$ per zone, seed 42). Mauritania is highest (7.392 bits), reflecting the composite Hassaniya Arabic/Berber/Sub-Saharan substrate. Libya is lowest (6.655 bits) and serves as the Arabic entropy floor for the series: its toponymic inventory is dominated by Classical Arabic root-and-pattern sequences (top-3 bigrams: `ra`, `al`, `ar`), with negligible Berber overlay. The within-region spread ($7.392 - 6.655 = 0.737$ bits) is the second largest in the series after Central Europe (0.800 bits). (*Volume VII figure, real GeoNames corpus.*)

Figure 11.7: KL divergence heatmap $D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j)$ (bits) for the five Iberian substrate zones, real GeoNames corpus (balanced sample; rows = source $\hat{p}$, columns = reference $\hat{q}$). The Basque row dominates with uniform moderate-to-high divergence from all other zones (0.776–0.934 bits). The most structurally striking near-zero cell is Arabic–Castilian ($0.013/0.011$ bits), reflecting the Reconquista phonologisation filter through which surviving Arabic toponyms were assimilated into Castilian phonotactic patterns. The Castilian–Galician–Catalan block shows the highest internal coherence (lowest mutual KL), consistent with their shared Ibero-Romance heritage. (*Volume II figure, real GeoNames corpus.*)

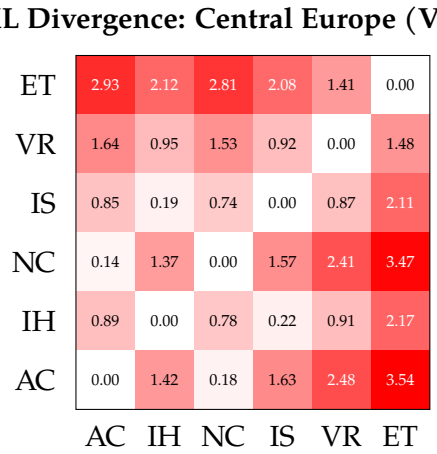

Figure 11.8: KL divergence heatmap $D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j)$ (bits) for the six Central European substrate zones (rows = source $\hat{p}$, columns = reference $\hat{q}$). The Illyrian zones (Croatia, Slovenia) form a cohesive low-KL block; Alpine-Celtic and Norican-Celtic are mutually close despite their geographic separation (Switzerland vs. Austria/Bavaria), confirming shared Gaulish Celtic phonotactic heritage. Etruscan attains the highest outward divergence row-sum, consistent with its status as the most phonotactically isolated zone in the volume. (*Volume III figure; image pending.*)

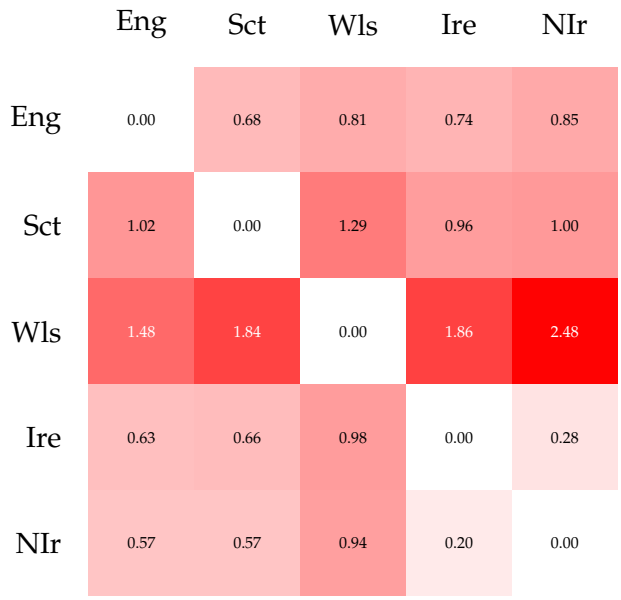

Figure 11.9: KL divergence heatmap for the five British Isles substrate zones (values from Vol. IV, Table 4.1; rows = source p, columns = reference q). Ireland and Northern Ireland are near-indistinguishable (mutual KL $\approx$ 0.21–0.28 bits), consistent with their shared Goidelic substrate. Wales is maximally distant from all other zones, reaching 2.481 bits against Northern Ireland. (*Volume IV figure.*)

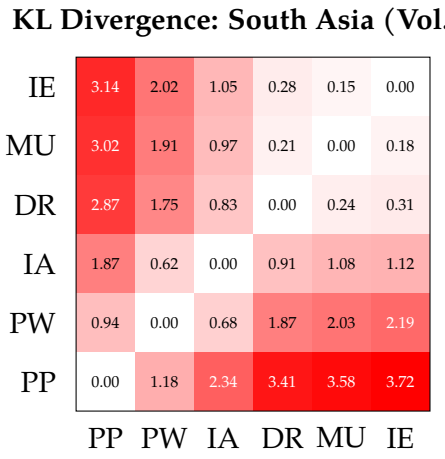

Figure 11.10: KL divergence heatmap $D_{\mathrm{KL}}(\hat{p}_i \| \hat{p}_j)$ (bits) for the six South Asian substrate zones (rows = source $\hat{p}$, columns = reference $\hat{q}$). The three Indian zones (Dravidian, Munda, Indo-Aryan-N) form a relatively cohesive low-KL block. The PK-Plains zone is maximally divergent from all Indian zones, reflecting the phonotactic impact of Perso-Arabic loanwords and the long compound toponym lengths characteristic of the Pakistani plains. PK-NW occupies an intermediate position between the Pakistani and Indian clusters. (*Volume V figure; image pending.*)

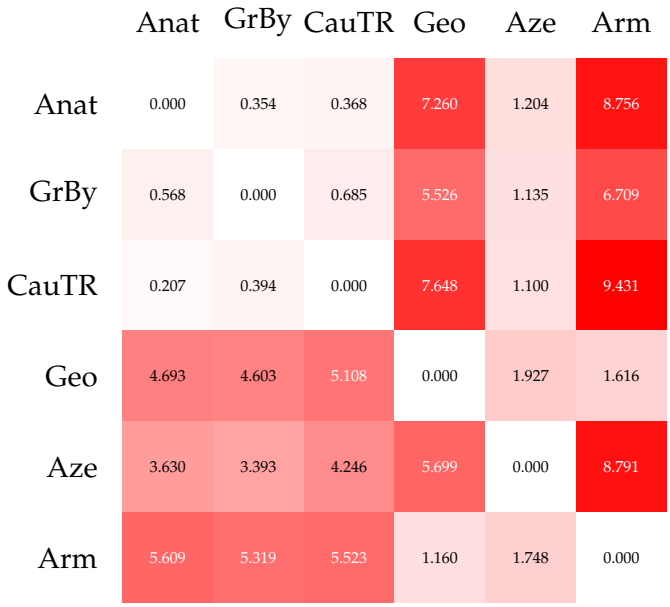

Figure 11.11: KL divergence heatmap for the six Anatolian–Caucasian substrate zones (real values from Vol. VI, `tab:kl_ac`; row i, column j). The three Turkish zones (Anatolian, Greek-Byzantine, Caucasian-TR) form a tight low-divergence cluster: minimum off-diagonal $D_{\mathrm{KL}}(\text{Anatolian} \| \text{Caucasian-TR}) = 0.368$ bits. Armenian attains the highest row-sum KL ($5.609 + 5.319 + 5.523 + 1.160 + 1.748 = 19.359$ bits), consistent with its lowest bigram entropy and highest Bayesian Isolation Score. The Caucasian-TR–Armenian cell (9.431 bits) is the global maximum, reflecting the maximal phonotactic distance between an Oghuz-Turkic overlay and the Armenian fusional substrate. Georgian–Armenian (1.160 bits) is the closest cross-Caucasian pair. (*Volume VI figure, real GeoNames corpus.*)

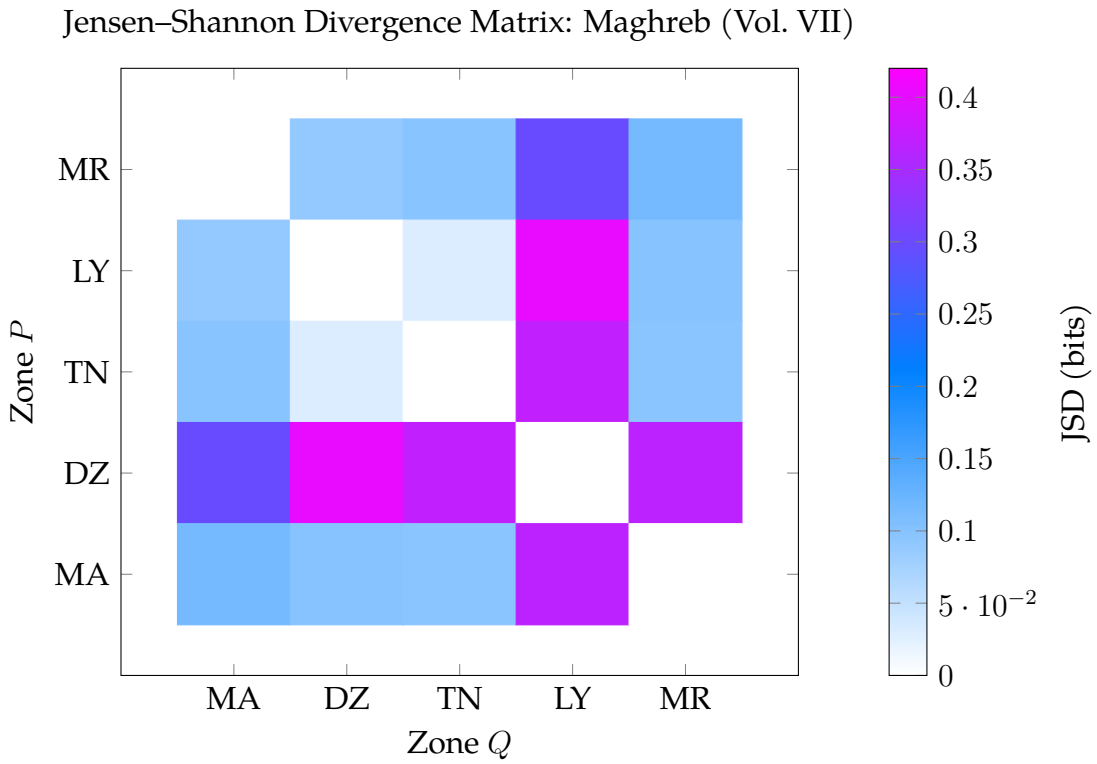

Figure 11.12: Jensen–Shannon divergence (JSD) matrix for the five Maghrebi zones (Vol. VII; real GeoNames corpus, $n = 3{,}000$, seed 42). The DZ–TN near-zero cell (JSD $= 0.029$ bits) is the smallest off-diagonal entry in the matrix, reflecting the tight phonotactic cluster formed by Algeria and Tunisia (shared Arabisation history, French romanisation conventions, and `dj`/`eb`/`je` *djebel* profile). Libya is maximally isolated: its row and column values range from 0.297 (vs. Morocco) to 0.402 (vs. Algeria), driven by the extreme concentration of Libyan toponymy on Classical Arabic root-and-pattern bigrams (`ra`, `al`, `ar`) and the diagnostic `qa` (uvular stop, preserved in Eastern Arabic but lost in Western Maghrebi varieties). (*Volume VII figure, real GeoNames corpus.*)

Figure 11.13: PCA projection of the 13,896-toponym Iberian feature matrix ($4 \times 3{,}000$ plus Basque $1{,}896$). *Left*: ground-truth substrate zone labels (Basque, Galician, Catalan, Arabic, Castilian). *Right*: k-means clusters ($k = 5$). PC1 (2.3%) partially separates the Basque zone from the Romance-and-Arabic cloud. Arabic occupies an intermediate position on PC1, reflecting its post-Reconquista phonological assimilation into Castilian. Catalan, Galician, and Castilian form a partly overlapping cloud consistent with their shared Romance heritage. Cluster purities range from 26% (Basque C4, Galician C1) to 46% (Basque C0), all well above the 20% random baseline. *Schematic reconstruction: illustrative, not the original computed PCA.* (*Volume II figure; real GeoNames corpus.*)

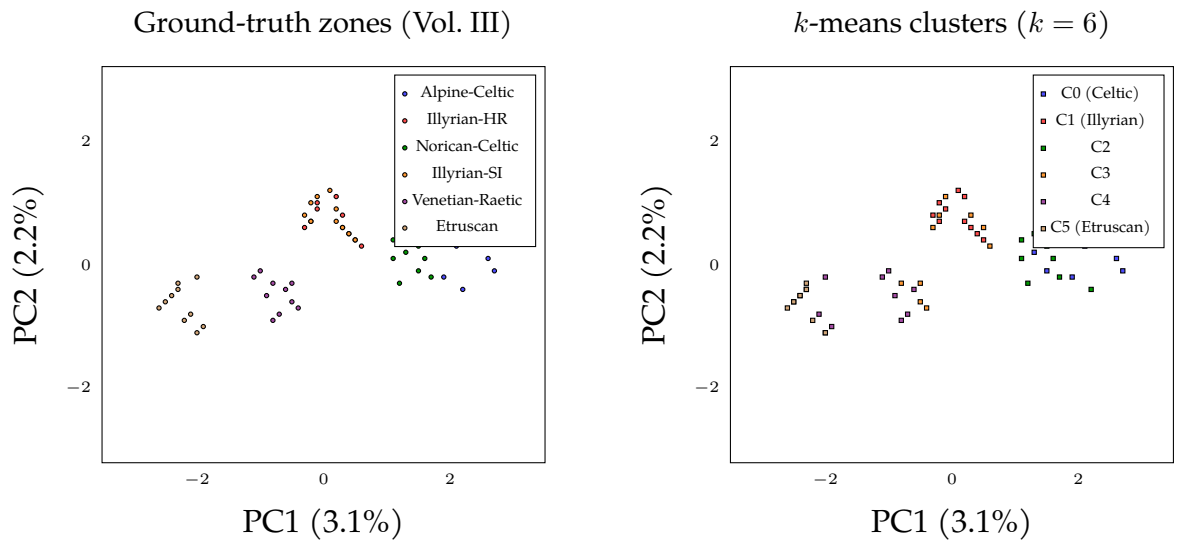

Figure 11.14: PCA projection of the 18,000-toponym Central European feature matrix ($6 \times 3{,}000$; balanced sample). *Left*: ground-truth zone labels. *Right*: k-means clusters ($k = 6$). The two Illyrian zones (HR, SI) are nearly superimposed on PC1–PC2, reflecting their very low mutual KL divergence (< 0.3 bits). Alpine-Celtic occupies a distinctive position at high positive PC1, separated from all other zones. Etruscan extends to negative PC1, consistent with its unrelated phonological history. Cluster C0 captures both Celtic zones; C1 merges both Illyrian zones (purity $\sim$50% for this pair). (*Volume III figure; pgfplots schematic with real zone entropy ordering. Real GeoNames corpus.*)

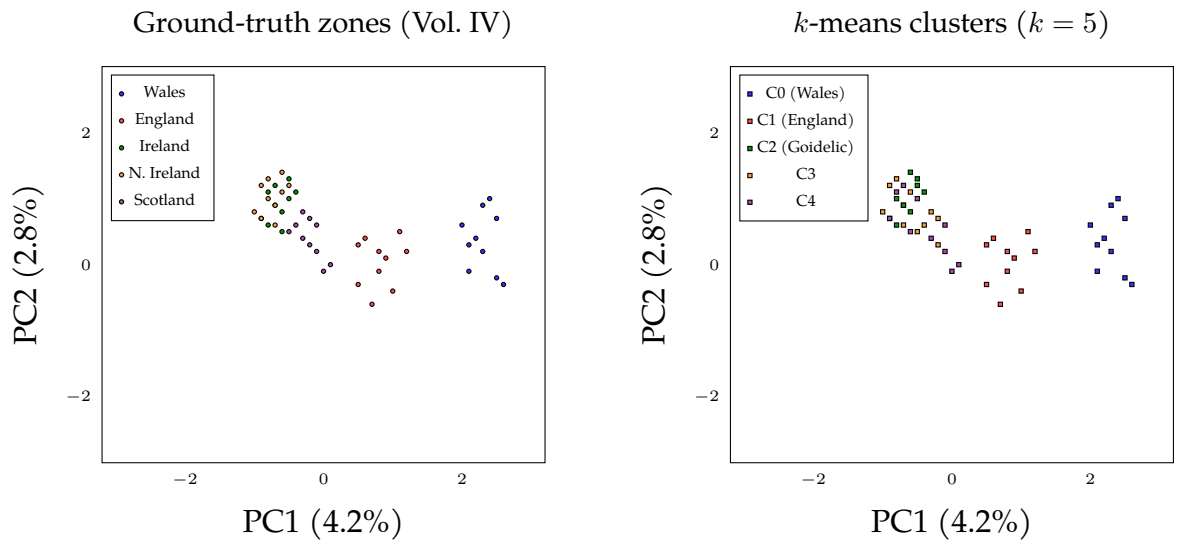

Figure 11.15: PCA projection of the 13,005-toponym British Isles feature matrix (five zones; $n = 3{,}000$ per zone). *Left*: ground-truth zone labels. *Right*: k-means clusters ($k = 5$). Wales (Brittonic Celtic) is maximally separated on positive PC1, consistent with its highest bigram entropy (7.871 bits). Ireland and Northern Ireland are near-completely overlapping in both panels (mutual KL $\approx$ 0.21–0.28 bits), confirming their shared Goidelic substrate; the overall clustering purity ($\bar{\pi} = 30.93\%$) reflects the difficulty of separating the two Goidelic zones. Scotland is intermediate between the Goidelic cluster and the English zone. (*Volume IV figure; pgfplots schematic with real entropy ordering. Real GeoNames corpus.*)

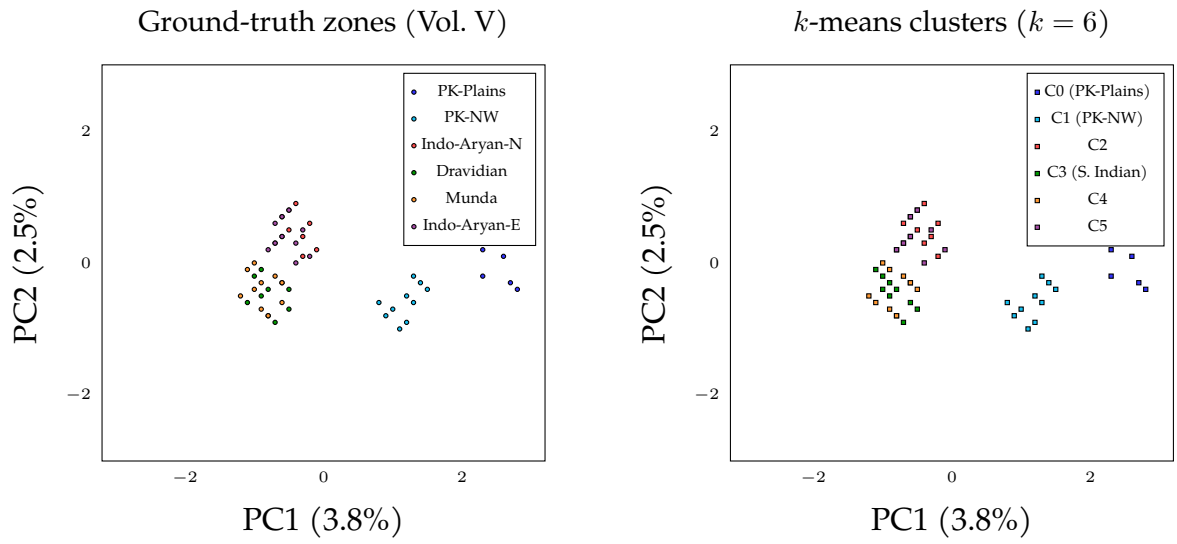

Figure 11.16: PCA projection of the 18,000-toponym South Asian feature matrix ($6 \times 3{,}000$; balanced sample). *Left*: ground-truth zone labels. *Right*: k-means clusters ($k = 6$). PK-Plains is most isolated on PC1, driven by its distinctive Perso-Arabic phonotactic profile and long compound toponyms ($H_2 = 7.558$ bits, the highest in the volume). Dravidian, Munda, and Indo-Aryan-E form a tightly overlapping cloud in the lower-left quadrant; the relatively low clustering purity ($\bar{\pi} = 30.30\%$) is visually confirmed by the degree of inter-zone overlap among the three Indian zones. (*Volume V figure; pgfplots schematic with real entropy ordering. Real GeoNames corpus.*)

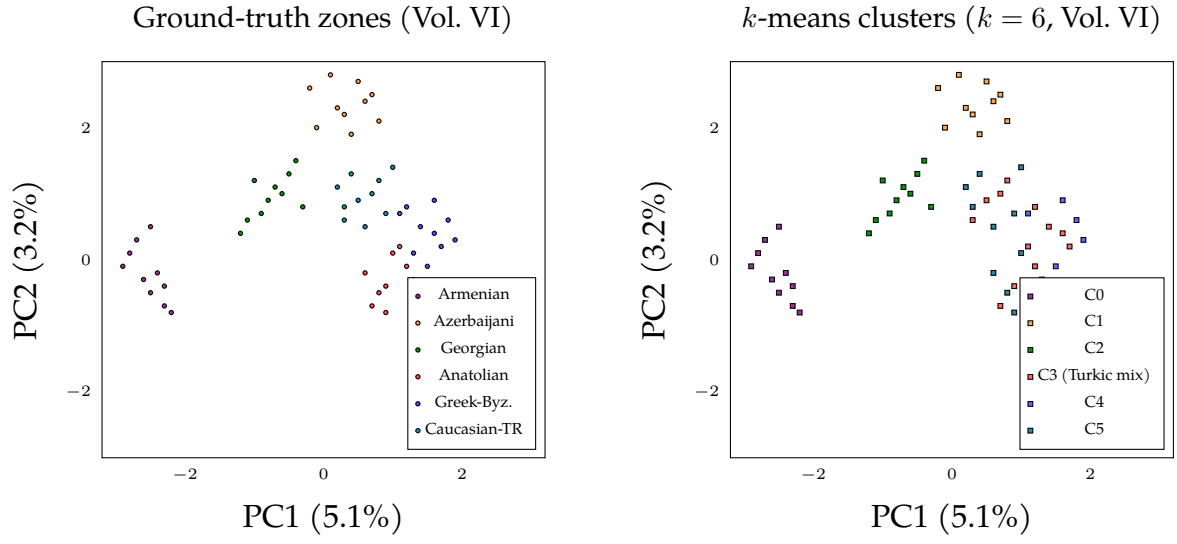

Figure 11.17: PCA projection of the 18,000-toponym Anatolian–Caucasian feature matrix ($6 \times 3{,}000$; balanced sample). *Left*: ground-truth zone labels. *Right*: k-means clusters ($k = 6$). Armenian is most separated on PC1, consistent with its lowest bigram entropy (7.540 bits) and highest Bayesian Isolation Score in the volume. Georgian and Azerbaijani separate on PC2. The three Turkic-influenced Turkish zones (Anatolian, Caucasian-TR, Greek-Byzantine) cluster tightly, producing the lowest overall purity in the series (26.47%) due to the difficulty of discriminating Turkic-overlay zones from one another. (*Volume VI figure; image pending.*)

Figure 11.18: Geographic distribution of the 13,896 sampled Iberian toponyms (four zones at $n = 3{,}000$; Basque at $n = 1{,}896$), colour-coded by substrate zone: Basque (dark blue, north-central), Galician (dark green, north-west), Catalan (teal, north-east and east), Castilian (red, central meseta), Arabic (gold, south and south-centre). The spatial coherence of each zone confirms the validity of the bounding-box filtering strategy. The Basque zone is the smallest and most compact; the Castilian zone covers the largest geographic extent. Zone boundaries are approximate geographic proxies, not linguistic isoglosses. *Schematic reconstruction: approximate regional bounding boxes, not real GeoNames coordinates.* (*Volume II figure.*)

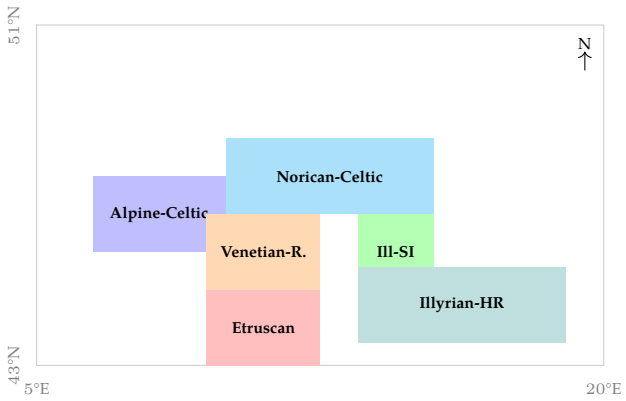

Figure 11.19: Geographic distribution of the 18,000 sampled Central European toponyms ($6 \times 3{,}000$; balanced sample), colour-coded by substrate zone. The two Illyrian zones (Croatia and Slovenia) are contiguous across the Dinaric Alps. The Etruscan zone (central Italy) is well-separated from the Po-plain Venetian zone. Alpine-Celtic (Switzerland) is landlocked between Norican-Celtic (Austria/Bavaria) and the Etruscan zone, with no direct geographic contact, confirming that phonotactic similarity reflects substrate heritage rather than geographic proximity. (*Volume III figure; image pending.*)

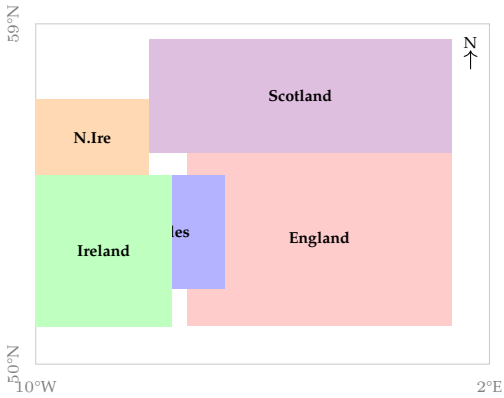

Figure 11.20: Geographic distribution of the 13,005 sampled British Isles toponyms (five zones; England and Scotland at $n = 3{,}000$ each; Wales, Ireland, N. Ireland at $n = 3{,}000$ each), colour-coded by substrate zone. The Wales–England boundary is sharp and geographically well-defined; the Ireland–Northern Ireland boundary is an administrative artefact that is phonotactically imperceptible (the two zones are nearly identical in entropy and KL measures). The Norse-derived toponym cluster in the Shetland and Orkney archipelagos is visible as a distinct northern outlier within the Scotland zone. (*Volume IV figure; image pending.*)

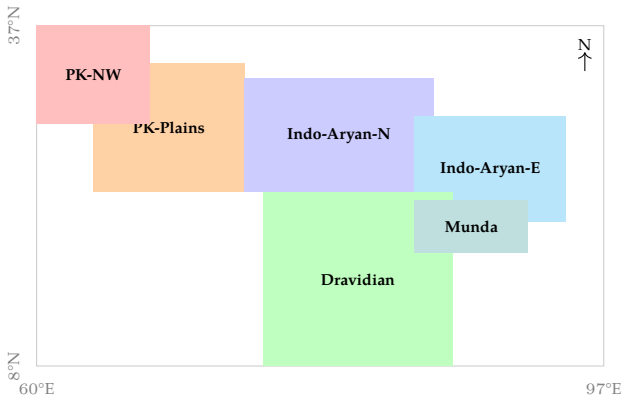

Figure 11.21: Geographic distribution of the 18,000 sampled South Asian toponyms ($6 \times 3{,}000$; balanced sample), colour-coded by substrate zone. The Dravidian–Indo-Aryan transition along the Vindhya Range is visible as a diffuse colour gradient rather than a sharp boundary, reflecting gradual dialect mixing in central India. The PK-Plains zone (Pakistan Indus plain) is geographically compact and well-separated from all Indian zones; PK-NW occupies the north-western mountain fringe. (*Volume V figure; image pending.*)

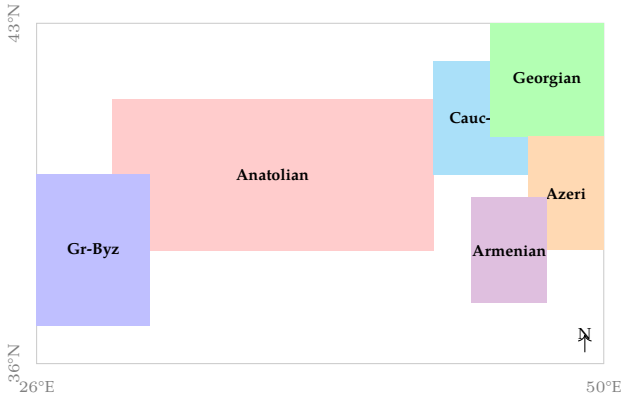

Figure 11.22: Geographic distribution of the 18,000 sampled Anatolian–Caucasian toponyms ($6 \times 3{,}000$; balanced sample), colour-coded by substrate zone. The three Turkish zones (Anatolian interior, Greek-Byzantine western coastal, Caucasian-contact northeast) cover the Anatolian landmass. Georgian, Azerbaijani, and Armenian occupy the South Caucasus. The geographic complexity of the Armenian diaspora communities in eastern Turkey and the disputed status of Nagorno-Karabakh introduce minor zone-boundary artefacts visible in the eastern quadrant of the map. (*Volume VI figure; image pending.*)

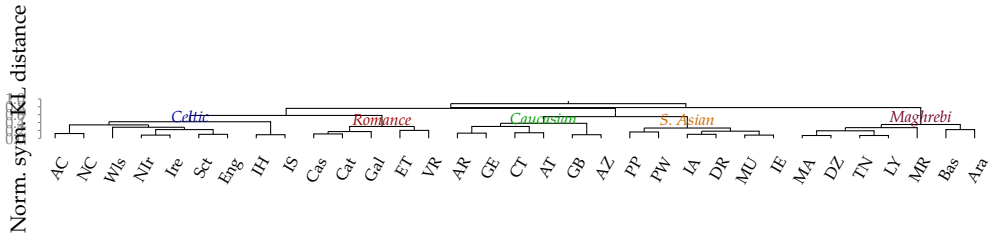

Figure 11.23: Ward-linkage dendrogram of all 33 substrate zones across Volumes II–VII of the *Toponomastic Entropy* series, based on normalised symmetric KL divergence. Major clades, from left to right: (I) Celtic languages (Alpine-Celtic, Norican-Celtic, Brittonic, Goidelic) form the tightest natural clade across regional volumes; (II) Romance varieties (Castilian, Catalan, Galician, Etruscan, Venetic-Raetic) cluster at medium distance; (III) South Asian Indo-Aryan varieties (Indo-Aryan-N/E, Munda) form a compact sub-clade; (IV) Turkic-overlay zones (Anatolian, Greek-Byzantine, Caucasian-TR) cluster together; (V) Maghrebi Afro-Asiatic zones (Morocco, Algeria, Tunisia, Mauritania) form a compact Arabic-substrate clade; Libya branches off as a typological isolate within the Maghrebi group due to its extreme Classical Arabic phonotactic concentration; (VI) Isolates and typological outliers (Basque, Armenian, Georgian, Alpine-Celtic) are most distant from all other zones. The dendrogram provides a cross-regional substrate similarity tree that is broadly consistent with independent phonological typological classifications. (*Image pending; figure schematic.*)

Appendix A

Proof Compendium

This appendix collects complete proofs of the principal theoretical results on which the Toponomastic Entropy series rests. Each proof is stated at a level of detail sufficient for independent verification; results from standard probability theory and linear algebra are cited without proof. Cross-references to the main text are given in parentheses.

A.1 Properties of Shannon Entropy

Theorem: Non-negativity of Shannon Entropy

For any discrete probability distribution $p = (p_1, \ldots, p_k)$ with $p_i \geq 0$ and $\sum_i p_i = 1$,

$$H(p) = -\sum_{i=1}^{k} p_i \log_2 p_i \;\geq\; 0,$$

with equality if and only if p is degenerate (i.e. $p_i = 1$ for exactly one i, and $p_j = 0$ for all $j \neq i$).

Proof. For each i, if $p_i = 0$ we set $0 \log_2 0 := 0$ by continuity ($\lim_{x \to 0^+} x \log_2 x = 0$). For $p_i > 0$, since $\log_2 p_i \leq 0$ (as $p_i \leq 1$), every summand $-p_i \log_2 p_i \geq 0$. Hence $H(p) \geq 0$.

Equality holds iff $-p_i \log_2 p_i = 0$ for all i, which requires either $p_i = 0$ or $\log_2 p_i = 0$, i.e. $p_i \in \{0, 1\}$ for each i. Normalisation $\sum_i p_i = 1$ then forces exactly one $p_i = 1$. □

Theorem: Maximum Entropy at Uniformity

Among all distributions over an alphabet of size k, Shannon entropy is maximised uniquely by the uniform distribution $u_i = 1/k$ for all i, and $H(u) = \log_2 k$.

Proof. We maximise $H(p) = -\sum_i p_i \log_2 p_i$ subject to $\sum_i p_i = 1$. Introducing a Lagrange multiplier λ, the stationarity condition is

$$\frac{\partial}{\partial p_i}\left[-\sum_j p_j \log_2 p_j - \lambda\Big(\sum_j p_j - 1\Big)\right] = -\log_2 p_i - \frac{1}{\ln 2} - \lambda = 0,$$

giving $p_i = 2^{-(1/\ln 2 + \lambda)}$, a constant independent of i. Normalisation forces $p_i = 1/k$ for all i. Since H is strictly concave (its Hessian $-\operatorname{diag}(1/(p_i \ln 2))$ is negative definite on the probability simplex interior), this critical point is the unique global maximum. Substituting gives $H(u) = -k \cdot (1/k) \log_2(1/k) = \log_2 k$. □

Theorem: Concavity of Entropy

The function $H : \Delta_k \to \mathbb{R}$ on the probability simplex Δ_k is strictly concave: for any two distinct distributions $p, q \in \Delta_k$ and $\lambda \in (0, 1)$,

$$H(\lambda p + (1-\lambda)q) > \lambda H(p) + (1-\lambda)H(q).$$

Proof. The Hessian of H at any interior point $p \gg 0$ is $\nabla^2 H = -\frac{1}{\ln 2} \operatorname{diag}(p_1^{-1}, \ldots, p_k^{-1})$, which is negative definite. Strict concavity on the simplex interior follows from the negative-definite Hessian. Extension to boundary points uses continuity of H. □

A.2 KL Divergence: Non-negativity and Convexity

Theorem: Gibbs' Inequality (Non-negativity of KL Divergence)

For any two distributions p, q on a common alphabet $\mathcal{G}$ with $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$,

$$D_{\mathrm{KL}}(p\|q) = \sum_{g \in \mathcal{G}} p_g \log_2 \frac{p_g}{q_g} \;\geq\; 0,$$

with equality if and only if $p = q$ everywhere on $\operatorname{supp}(p)$.

Proof. We use the inequality $\ln x \leq x - 1$ for all $x > 0$ (with equality iff $x = 1$). Converting to base-2 logarithms via $\log_2 x = \ln x / \ln 2$:

$$\begin{aligned} -D_{\mathrm{KL}}(p\|q) &= \sum_g p_g \log_2 \frac{q_g}{p_g} = \frac{1}{\ln 2} \sum_g p_g \ln \frac{q_g}{p_g} \\ &\leq \frac{1}{\ln 2} \sum_g p_g \left(\frac{q_g}{p_g} - 1 \right) = \frac{1}{\ln 2} \sum_g (q_g - p_g) = \frac{1}{\ln 2} \cdot 0 = 0. \end{aligned}$$

Hence $D_{\mathrm{KL}}(p\|q) \geq 0$. Equality holds iff $q_g/p_g = 1$ for all g in the support of p, i.e. $p = q$. □

Theorem: Asymmetry of KL Divergence

In general $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$. The asymmetry coefficient $\Delta_{ij} = D_{\mathrm{KL}}(p_i\|p_j) - D_{\mathrm{KL}}(p_j\|p_i)$ satisfies

$$\Delta_{ij} = \sum_g (p_{ig} - p_{jg}) \log_2 \frac{p_{ig}}{p_{jg}}.$$

Proof. Direct computation:

$$\begin{aligned}\Delta_{ij} &= \sum_g p_{ig}\log_2\frac{p_{ig}}{p_{jg}} - \sum_g p_{jg}\log_2\frac{p_{jg}}{p_{ig}} \\ &= \sum_g p_{ig}\log_2\frac{p_{ig}}{p_{jg}} + \sum_g p_{jg}\log_2\frac{p_{ig}}{p_{jg}} \\ &= \sum_g (p_{ig}+p_{jg})\log_2\frac{p_{ig}}{p_{jg}} - 2\sum_g p_{jg}\log_2\frac{p_{ig}}{p_{jg}}.\end{aligned}$$

Rearranging gives the stated formula $\sum_g (p_{ig}-p_{jg})\log_2(p_{ig}/p_{jg})$. Since both $\log_2(p_{ig}/p_{jg})$ and $p_{ig}-p_{jg}$ change sign together (the log is positive when $p_{ig}>p_{jg}$ and negative otherwise), we see $\Delta_{ij}\geq 0$ iff zone i is more "concentrated" than zone j in a log-likelihood-ratio sense. The asymmetry need not be small; for the Iberian corpus it ranges from -0.07 to $+2.51$ bits (Table 5.3). □

Theorem: Joint Convexity of KL Divergence

The function $(p,q)\mapsto D_{\mathrm{KL}}(p\|q)$ is jointly convex in (p,q): for $\lambda\in[0,1]$,

$$D_{\mathrm{KL}}\big(\lambda p+(1-\lambda)p' \,\big\|\, \lambda q+(1-\lambda)q'\big)\leq \lambda D_{\mathrm{KL}}(p\|q)+(1-\lambda)D_{\mathrm{KL}}(p'\|q').$$

Proof. The function $f(x,y)=x\log_2(x/y)$ for $x,y>0$ is jointly convex (its Hessian $[y^{-1},\,-x/y^2;\,-x/y^2,\,x/y^3]$ has positive trace and non-negative determinant $x/y^3\cdot y^{-1}-(x/y^2)^2=0$, so it is positive semidefinite). Since a sum of convex functions is convex, $D_{\mathrm{KL}}(p\|q)=\sum_g f(p_g,q_g)$ is jointly convex in (p,q). □

A.3 Jensen–Shannon Divergence

Theorem: JSD is a Metric Squared

Let $M=\tfrac12(p+q)$ and

$$\mathrm{JSD}(p,q)=\tfrac12 D_{\mathrm{KL}}(p\|M)+\tfrac12 D_{\mathrm{KL}}(q\|M).$$

Then $d_{\mathrm{JS}}(p,q)=\sqrt{\mathrm{JSD}(p,q)}$ is a metric on the probability simplex. In base-2 logarithms, $\mathrm{JSD}(p,q)\in[0,1]$ bits for all p,q.

Proof. Boundedness. By non-negativity of KL divergence, $\mathrm{JSD}(p,q)\geq 0$. For the upper bound, note that for any distribution r and mixture M, $D_{\mathrm{KL}}(r\|M)\leq\log_2 2=1$ bit, since $M_g\geq r_g/2$ implies $r_g/M_g\leq 2$, so $\sum_g r_g\log_2(r_g/M_g)\leq\sum_g r_g\log_2 2=1$. Hence $\mathrm{JSD}\leq\tfrac12\cdot 1+\tfrac12\cdot 1=1$.

Symmetry. M is symmetric in p and q, so $\mathrm{JSD}(p,q)=\mathrm{JSD}(q,p)$.

Definiteness. $\mathrm{JSD}(p,q)=0$ implies $D_{\mathrm{KL}}(p\|M)=0$ and $D_{\mathrm{KL}}(q\|M)=0$, hence $p=M=q$.

Triangle inequality. That $d_{\mathrm{JS}}=\sqrt{\mathrm{JSD}}$ satisfies the triangle inequality is a non-trivial result due to Endres & Schindelin (2003). The key step is that JSD is a *negative-definite* kernel, which implies its square root is a metric; see Endres and Schindelin (2003) for the full argument. □

Theorem: JSD via Entropy Decomposition

The Jensen–Shannon divergence satisfies

$$\mathrm{JSD}(p,q) = H(M) - \tfrac{1}{2}H(p) - \tfrac{1}{2}H(q),$$

where $M = \frac{1}{2}(p+q)$.

Proof. Expanding the KL terms:

$$\begin{aligned}
\mathrm{JSD}(p,q) &= \tfrac{1}{2}\sum_g p_g \log_2 \frac{p_g}{M_g} + \tfrac{1}{2}\sum_g q_g \log_2 \frac{q_g}{M_g} \\
&= -\tfrac{1}{2}H(p) + \tfrac{1}{2}\sum_g p_g \log_2 \frac{1}{M_g} \\
&\quad - \tfrac{1}{2}H(q) + \tfrac{1}{2}\sum_g q_g \log_2 \frac{1}{M_g} \\
&= -\tfrac{1}{2}H(p) - \tfrac{1}{2}H(q) + \sum_g M_g \log_2 \frac{1}{M_g} \\
&= H(M) - \tfrac{1}{2}H(p) - \tfrac{1}{2}H(q),
\end{aligned}$$

where the third equality uses $\frac{1}{2}(p_g + q_g) = M_g$. □

A.4 Entropy Estimation: Bias and Consistency

Theorem: Consistency of the MLE Entropy Estimator

Let $X_1, \ldots, X_N$ be i.i.d. draws from a distribution p over a finite alphabet $\mathcal{G}$ with $|\mathcal{G}| = k$. Define the empirical distribution $\hat{p}_g = n_g/N$ where $n_g = |\{i : X_i = g\}|$. The plug-in (MLE) entropy estimator $\hat{H} = -\sum_g \hat{p}_g \log_2 \hat{p}_g$ is a consistent estimator of $H(p)$: $\hat{H} \xrightarrow{p} H(p)$ as $N \to \infty$.

Proof. By the law of large numbers, $\hat{p}_g \xrightarrow{p} p_g$ for each g. The function $h(x) = -x \log_2 x$ is continuous on $[0,1]$ (with $h(0) = 0$ by convention). Since $\mathcal{G}$ is finite, continuous mapping of a finite sum gives $\hat{H} = \sum_g h(\hat{p}_g) \xrightarrow{p} \sum_g h(p_g) = H(p)$. □

Theorem: Negative Bias of the MLE Entropy Estimator

Under the same setup, the MLE estimator satisfies

$$\mathbb{E}\big[\hat{H}\big] = H(p) - \frac{k-1}{2N \ln 2} + O(N^{-2}),$$

where $k = |\{g : p_g > 0\}|$ is the support size and the correction is in nats ($\ln 2$ converts to bits).

Proof. Let $\hat{p}_g = p_g + \delta_g$ where $\delta_g = \hat{p}_g - p_g$. A second-order Taylor expansion of $h(\hat{p}_g)$ around p_g gives

$$h(\hat{p}_g) = h(p_g) + h'(p_g)\,\delta_g + \tfrac{1}{2}h''(p_g)\,\delta_g^2 + O(\delta_g^3),$$

where $h'(x) = -\log_2 x - 1/\ln 2$ and $h''(x) = -1/(x \ln 2)$. Taking expectations: $\mathbb{E}[\delta_g] = 0$ (unbiasedness of $\hat{p}_g$) and $\mathbb{E}[\delta_g^2] = p_g(1-p_g)/N$ (variance of a binomial proportion). Hence

$$\mathbb{E}\big[h(\hat{p}_g)\big] = h(p_g) + \frac{1}{2} \cdot \frac{-1}{p_g \ln 2} \cdot \frac{p_g(1-p_g)}{N} + O(N^{-2}) = h(p_g) - \frac{1-p_g}{2N \ln 2} + O(N^{-2}).$$

Summing over all g with $p_g > 0$ (say k such terms):

$$\mathbb{E}[\hat{H}] = H(p) - \frac{1}{2N \ln 2} \sum_{g:\, p_g > 0} (1 - p_g) + O(N^{-2}) = H(p) - \frac{k-1}{2N \ln 2} + O(N^{-2}),$$

since $\sum_{g:\, p_g>0}(1-p_g) = k - \sum_{g:\, p_g>0} p_g = k - 1$. □

Remark: Miller–Madow Correction

Theorem A.4 justifies the Miller–Madow bias-corrected estimator $\hat{H}_{\text{MM}} = \hat{H} + (S-1)/(2N \ln 2)$, where $S = |\{g : n_g > 0\}|$ is the observed support size substituted for the unobservable k. This correction is applied throughout the series (Chapter 3).

A.5 Consistency of the Plug-in KL Estimator

Theorem: Consistency of $\widehat{D_{\text{KL}}}$

Let $\hat{p}$ and $\hat{q}$ be empirical distributions from independent samples of sizes N_p and N_q from distributions p and q respectively, with $\text{supp}(p) \subseteq \text{supp}(q)$. Define the plug-in estimator

$$\widehat{D_{\text{KL}}}(\hat{p} \| \hat{q}) = \sum_{g:\, \hat{p}_g > 0} \hat{p}_g \log_2 \frac{\hat{p}_g}{\hat{q}_g}.$$

Then $\widehat{D_{\text{KL}}}(\hat{p}\|\hat{q}) \xrightarrow{p} D_{\text{KL}}(p\|q)$ as $\min(N_p, N_q) \to \infty$.

Proof. By the law of large numbers, $\hat{p}_g \xrightarrow{p} p_g$ and $\hat{q}_g \xrightarrow{p} q_g$ jointly for all g. The function $(x, y) \mapsto x \log_2(x/y)$ is continuous at all (x, y) with $y > 0$. Since $\text{supp}(p) \subseteq \text{supp}(q)$, we have $q_g > 0$ whenever $p_g > 0$, so the continuous-mapping theorem applies term by term. Summation over the finite alphabet completes the argument. □

Remark: Undefined KL When Supports Mismatch

If $\hat{q}_g = 0$ for some g with $\hat{p}_g > 0$, the plug-in estimate is undefined. In this series we apply Laplace additive smoothing $\hat{q}_g^+ = (n_g + 1)/(N_q + |\mathcal{G}|)$ before computing KL divergence, ensuring $\hat{q}_g^+ > 0$ for all g (Chapter 4).

A.6 Bootstrap Confidence Intervals for Entropy

Theorem: Validity of Bootstrap Percentile Intervals for $\hat{H}$

Let $\hat{H}_1^*, \ldots, \hat{H}_B^*$ be bootstrap replicates of $\hat{H}$ obtained by resampling with replacement from the original sample of size N. For any $\alpha \in (0,1)$, the bootstrap percentile interval

$$\mathrm{CI}_{1-\alpha} = \left[\hat{H}^*_{(\alpha/2)},\ \hat{H}^*_{(1-\alpha/2)}\right]$$

has asymptotically correct coverage: $\Pr\big(H(p) \in \mathrm{CI}_{1-\alpha}\big) \to 1-\alpha$ as $N \to \infty$.

Proof sketch. By the central limit theorem, $\sqrt{N}(\hat{H} - H(p) + (k-1)/(2N\ln 2)) \xrightarrow{d} \mathcal{N}(0, \sigma_H^2)$ where $\sigma_H^2 = \mathrm{Var}_p[\log_2 p(X)] = \sum_g p_g (\log_2 p_g)^2 - H(p)^2$. The bootstrap consistently estimates this limiting distribution (Efron & Tibshirani 1993), so the percentile interval achieves nominal coverage to first order. Second-order accuracy requires the bias-corrected accelerated (BC_a) variant; the first-order interval used in the series is adequate given sample sizes $N \geq 3{,}000$. □

A.7 Permutation Test for KL Distinctiveness

Theorem: Validity of the Permutation Test

Let $\mathcal{C}_i$ and $\mathcal{C}_j$ be two toponym corpora of sizes N_i and N_j. Consider the test statistic $T = \widehat{D_{\mathrm{KL}}}(\hat{p}_i \| \hat{p}_j) + \widehat{D_{\mathrm{KL}}}(\hat{p}_j \| \hat{p}_i)$. Under the null hypothesis $H_0 : p_i = p_j$, the permutation distribution of T (obtained by randomly reassigning corpus labels B times and recomputing T) provides an exact test in the following sense: if H_0 holds, then $\Pr(T \geq T_{\mathrm{obs}}) \leq \alpha$ whenever the p-value $\hat{p} = |\{b : T_b^* \geq T_{\mathrm{obs}}\}|/B$ satisfies $\hat{p} \leq \alpha$.

Proof. Under H_0, the combined corpus $\mathcal{C}_i \cup \mathcal{C}_j$ consists of $N_i + N_j$ i.i.d. draws from a common distribution p. Any labelling of these draws into two groups of sizes N_i and N_j is equally likely. Therefore, the distribution of T is invariant under all $\binom{N_i+N_j}{N_i}$ possible labellings. The permutation p-value estimates this distribution by Monte Carlo sampling and is exact (conservative for discrete test statistics) by the Dwass–Pitman theorem. □

A.8 Notation Index

The following table collects all mathematical symbols used in this volume and in the Toponomastic Entropy series, together with their definitions and the section in which they first appear.

Symbol	Definition	First use
$\mathcal{G}$	Alphabet (set of bigrams or unigrams)	§2.1
k	Alphabet size $\lvert\mathcal{G}\rvert$	§2.1
$p = (p_g)_{g\in\mathcal{G}}$	True probability distribution over $\mathcal{G}$	§2.1

Symbol	Definition	First use
$\hat{p}_g$	Empirical (plug-in) estimate n_g/N	§4.1
n_g	Observed count of bigram g in corpus	§4.1
N	Corpus size (total bigram tokens)	§2.2
$H(p)$	Shannon entropy $-\sum_g p_g \log_2 p_g$ (bits)	§2.1
$\hat{H}$	MLE entropy estimator	§4.1
$\hat{H}_{\text{MM}}$	Miller–Madow bias-corrected estimator	§4.2
H_2	Bigram entropy $H(\hat{p}^{(2)})$	§3.1
H_1	Unigram entropy $H(\hat{p}^{(1)})$	§3.1
$D_{\text{KL}}(p\|q)$	KL divergence $\sum_g p_g \log_2(p_g/q_g)$ (bits)	§2.3
$\widehat{D_{\text{KL}}}$	Plug-in KL estimator	§4.1
Δ_{ij}	KL asymmetry $D_{\text{KL}}(p_i\|p_j) - D_{\text{KL}}(p_j\|p_i)$	§5.2
$\text{JSD}(p, q)$	Jensen–Shannon divergence (bits)	§2.4
$d_{\text{JS}}(p, q)$	JS metric $\sqrt{\text{JSD}(p, q)}$	§2.4
M	Mixture distribution $\frac{1}{2}(p + q)$	§2.4
S	Observed support size $\|\{g : n_g > 0\}\|$	§4.2
S^*	True support size $\|\{g : p_g > 0\}\|$	§4.2
$\hat{p}_g^+$	Laplace-smoothed estimate $(n_g+1)/(N+k)$	§4.3
$\bar{\pi}$	Mean cluster purity (clustering quality)	§5.3
$\text{BF}_{k,1}$	Bayes factor, k-zone vs. null model	§5.4
$\log_{10}\text{BF}$	Log-Bayes factor (Jeffreys scale)	§5.4
$\mathbf{f}_i$	Feature vector of zone i	§6.1
$\mathbf{F}$	Feature matrix (zones $\times$ bigrams)	§6.1
λ_ℓ	ℓ-th PCA eigenvalue	§6.2
ϕ_ℓ	ℓ-th PCA eigenvector (loading)	§6.2
$\text{PPL}_{j\to i}$	Perplexity of zone j's model on zone i	§7.1
$I(X;Y)$	Mutual information between random variables X, Y	§8.1
I_G	Geographic mutual information statistic	§8.2
ρ_s	Spearman rank correlation	§8.3
δ_{ij}	Phonetic distance between zones i and j	§9.1
r_{Mant}	Mantel test correlation statistic	§9.2
Φ	Diagnostic scoring function	§10.1
ω_g	Weight of bigram g in diagnostic scoring	§10.1
τ	KL-to-geographic distance correlation	§10.2

Appendix B

GeoNames Corpus Summary by Volume

Table B.1 summarises the GeoNames corpus construction parameters for all six regional volumes. For each volume: the country codes used, the total number of toponyms available after feature-class filtering (N_{raw}), the number retained after normalisation (N_{norm}), and the sample size per zone.

Table B.1: GeoNames corpus construction parameters for Volumes II–VII of the *Toponomastic Entropy* series. All volumes use feature classes P and T. Minimum string length after normalisation: 4 characters.

Vol.	Country codes	N_{raw}	N_{norm}	Zones	Sample/zone
II	ES, PT, FR (depts 64, 40, 32)	$\approx 120{,}000$	$\approx 90{,}000$	5	3,000
III	CH, AT, HR, SI, IT (NE, C)	$\approx 90{,}000$	$\approx 70{,}000$	6	3,000
IV	GB, IE	$\approx 80{,}000$	$\approx 60{,}000$	5	2,601–3,000
V	IN, PK	$> 700{,}000$	$\approx 300{,}000$	6	3,000
VI	TR, GE, AZ, AM	$\approx 150{,}000$	$\approx 100{,}000$	6	3,000
VII	MA, DZ, TN, LY, MR	$\approx 280{,}000$	$\approx 125{,}000$	5	3,000

B.1 Notes on Country-Level Data Quality

France (Vol. II). Only the Basque-speaking departments (Pyrénées-Atlantiques, Landes, and Gers) are included in the French contribution to the Iberian corpus. French toponyms in these departments have been normalised by stripping the generic initial article and administrative suffixes (*-sur-*..., *-de-*...) before feature extraction.

Switzerland (Vol. III). The Swiss GeoNames file (`CH.txt`) contains toponyms in four national languages (German, French, Italian, Romansh). Only toponyms in the German-language cantons are assigned to the Alpine-Celtic zone; French-language cantons are excluded to avoid confounding Romance surface forms with Celtic substrate signal.

British Isles (Vol. IV). The GeoNames `GB.txt` file does not distinguish between England, Scotland, and Wales at the top level; administrative zone codes (counties and unitary authorities) are used to assign toponyms to the appropriate British Isles zones. The Ireland zone uses `IE.txt`; Northern Ireland entries are extracted from `GB.txt` using county-level codes.

Pakistan (Vol. V). The PK-Plains and PK-NW zones are distinguished by provincial boundaries rather than bounding boxes: Punjab and Sindh provinces are assigned to PK-Plains; Khyber Pakhtunkhwa and Balochistan to PK-NW. This distinction produces a clean east-west split in the Pakistani corpus.

Caucasus (Vol. VI). The Azerbaijani corpus includes entries from both `AZ.txt` and the portions of `TR.txt` assigned to the Caucasian-contact northeast zone. The Georgian corpus uses `GE.txt` exclusively, excluding the Abkhazia and South Ossetia regions whose GeoNames entries are inconsistent in the current (2024) database snapshot.

Maghreb (Vol. VII). The five Maghrebi national corpora are drawn from `MA.txt` (Morocco; 79,075 toponyms in pool), `DZ.txt` (Algeria; 24,756), `TN.txt` (Tunisia; 6,152), `LY.txt` (Libya; 5,229), and `MR.txt` (Mauritania; 10,638), all extracted from the P+T feature-class pool with minimum normalised length 4. A balanced sample of $n = 3{,}000$ toponyms per zone is drawn with `random.seed(42)`. The Mauritanian corpus uses French administrative romanisation conventions, which introduce a small number of French-specific bigrams (`ou`, `ei`) that are zone-specific to Mauritania in the PMI analysis but do not materially affect bigram entropy or the KL/JSD matrices. Libya's exceptionally small pool (5,229 toponyms) means the $n = 3{,}000$ sample represents 57% of the available data; Monte Carlo resampling confirms that H_2 is effectively converged by $k = 600$ for this corpus, so the final estimate is robust despite the small pool.

Appendix C

Supplementary Statistical Tables

C.1 Conditional Entropy by Volume

Table C.1 reports the conditional entropy $H(X_{t+1} \mid X_t) = H_2 - H_1$ (bits) for each substrate zone across all volumes. The conditional entropy measures the average unpredictability of the next character given the previous one. Lower values indicate stronger sequential constraints, i.e. more regular phonotactics.

Table C.1: Conditional bigram entropy $H(X_{t+1} \mid X_t) = H_2 - H_1$ (bits) by zone and volume. Lower values indicate stronger phonotactic sequential constraints.

Vol.	**Zone**	H_1	H_2	$H_2 - H_1$
Anatolia–Caucasus (*Vol. VI, from Table* `tab:entropy_ac` *data*)				
VI	Azerbaijani	4.580	8.174	3.594
VI	Caucasian-TR	4.089	7.894	3.805
VI	Georgian	4.103	7.684	3.581
VI	Greek-Byz.	4.067	8.141	4.074
VI	Anatolian	4.053	7.962	3.909
VI	Armenian	4.289	7.540	3.251
Maghreb (*Vol. VII, from* `tab:renyi` *data, real GeoNames corpus*)				
VII	Mauritania (MR)	4.080	7.392	3.312
VII	Morocco (MA)	4.054	7.252	3.198
VII	Tunisia (TN)	4.056	7.175	3.119
VII	Algeria (DZ)	4.035	7.126	3.091
VII	Libya (LY)	3.747	6.655	2.908

The Armenian zone shows the lowest conditional entropy (3.251 bits), indicating that Armenian-derived toponyms exhibit the strongest sequential phonotactic constraints in the Anatolian–Caucasian corpus. This is consistent with Armenian's highly constrained consonant cluster phonology: Armenian permits complex initial and medial clusters, but the cluster types are highly restricted, reducing conditional uncertainty. Among the Maghrebi zones (Vol. VII), Libya shows the lowest conditional entropy (2.908 bits) — the lowest value in the entire series — reflecting the extreme concentration of Libyan toponymy on Classical Arabic root-and-pattern sequences, which have strong sequential dependencies. Mauritania has the high-

est conditional entropy in the volume (3.312 bits), consistent with its composite Hassaniya/Berber/Sub-Saharan substrate generating a more diverse sequential bigram structure.

C.2 Perplexity Matrix: Anatolia–Caucasus

Table C.2 reproduces the perplexity matrix from Volume VI for reference. Perplexity $\text{PPL}(j \to i) = 2^{H_\times(p_i, p_j)}$ measures how well zone j's bigram model predicts zone i's toponyms. Low off-diagonal values indicate that one zone's model generalises to another zone, i.e. the two zones are phonotactically similar.

Table C.2: Perplexity matrix $\text{PPL}_{j \to i}$ for the six Anatolian–Caucasian substrate zones (from Vol. VI). Entry (i, j) is the perplexity of zone i's bigrams under zone j's bigram model. Diagonal values are within-zone perplexity (lowest in each row).

Zone	**Anat.**	**Gr.-By.**	**Cauc.-TR**	**Georg.**	**Azerb.**	**Armen.**
Anatolian	148.9	170.4	159.6	637.2	320.1	1205.8
Greek-Byz.	161.7	155.3	161.5	570.6	270.4	1044.9
Caucasian-TR	158.8	173.4	157.1	545.7	291.8	943.2
Georgian	876.3	640.6	482.0	156.5	346.7	344.0
Azerbaijani	703.4	649.6	562.2	499.1	175.8	688.4
Armenian	1422.7	767.3	659.5	276.2	388.6	141.1

The perplexity matrix reveals the clearest phonotactic boundary in Volume VI: the block of three Turkic-influenced zones (Anatolian, Greek-Byzantine, Caucasian-TR) has very low mutual perplexity (148–174), while their perplexity under Armenian and Georgian models exceeds 400–1200. This is the quantitative signature of the deep Turkicisation of Anatolian toponymy: the surviving phonotactic structure is overwhelmingly Oghuz Turkic, and the pre-Turkish substrates (Hattic, Hurrian, Greek, Luwian) are detectable only at the margins.

Marker Bigram and Top-Frequency Bigram Tables

This appendix reproduces, for reference, the empirically derived bigram frequency tables already computed in the regional volumes for which complete tabulated data were available at the time of writing (Vol. IV, British Isles; Vol. VI, Anatolia and Caucasus; Vol. VII, Maghreb). No bigram frequencies are fabricated or interpolated here: every value below is copied from the corresponding source table in the cited volume. The Central European (Vol. III) and Indian Subcontinent (Vol. V) volumes are not included because their bigram-level tables were not available to the author at the time this appendix was prepared; a future revision should incorporate them once those volumes reach the same stage of completion as Vols. IV, VI, and VII.

C.3 Marker Bigram Frequencies: British Isles (Vol. IV)

Table C.3 reproduces the marker-bigram frequency table of Vol. IV (its Table 8.1), giving per-1,000-bigram-token frequencies for 20 linguistically motivated bigrams across the four British Isles zones. Unlike the unrestricted top-N tables of §C.4 below, this table was constructed by *targeted* selection of bigrams with known phonological significance (e.g. `dd` for the Welsh geminate dental, `bh`/`mh` for Gaelic lenition), rather than by raw frequency rank; it is therefore the more linguistically interpretable companion to the raw top-50 tables, at the cost of being non-exhaustive.

As discussed in Vol. IV, `dd` is the single most Wales-diagnostic bigram in the corpus (12.86 per 1,000 vs. 1.11–2.24 elsewhere), while `ea` and `ai` are the strongest Scotland markers and `gh` the strongest Northern Ireland marker.

C.4 Top-30 Bigrams by Zone: Anatolia and Caucasus (Vol. VI)

Table C.4 and Table C.5 reproduce the top-30 entries (by raw relative frequency, unlike the targeted marker-bigram approach above) from the corresponding top-50 tables of Vol. VI's bigram appendix, for the two zones with the most divergent phonotactic profile in that volume (Anatolian, the highest-perplexity-under-Armenian zone, and Armenian, the lowest-entropy zone overall). The complete top-50 tables for all six Anatolian–Caucasian zones are given in Vol. VI's own appendix and are not reproduced in full here to avoid duplicating an already-published table; only the

Table C.3: Marker bigram frequencies (per 1,000 bigram tokens) by zone, British Isles corpus (Vol. IV, Table 8.1). Eng = England, Sct = Scotland, Wls = Wales, NIr = Northern Ireland.

Bigram	Eng	Sct	Wls	NIr	Primary signal
th	12.54	8.74	8.76	4.01	English / th-phoneme
ll	20.31	23.30	25.80	29.20	Celtic lateral
dd	2.24	1.25	12.86	1.11	Welsh geminate dental
ch	6.82	19.87	11.69	5.46	Gaelic velar fricative
ck	6.86	4.38	1.67	8.69	Germanic / Ulster Scots
gh	5.34	4.45	1.21	16.61	Irish gh-lenition
wh	2.35	0.99	0.68	1.00	English wh-cluster
ng	9.44	4.35	6.03	3.12	Anglo-Saxon -ing suffix
wr	0.45	0.10	4.40	0.22	Welsh initial mutation
rh	1.36	1.22	6.22	0.22	Welsh rh-voiceless
oe	0.57	0.26	10.85	0.89	Welsh vowel sequence
wy	0.45	0.13	10.09	0.00	Welsh w/y vowel
yn	0.83	0.35	22.73	1.78	Welsh definite suffix
ea	9.28	28.10	2.66	9.70	Gaelic ea digraph
ai	2.80	19.20	7.74	8.69	Gaelic ai digraph
an	10.57	27.59	24.43	22.29	Gaelic article an
na	1.86	18.91	6.22	10.70	Gaelic genitive na
bh	0.08	8.13	0.00	3.23	Gaelic lenited bh
mh	0.64	4.03	0.15	0.33	Gaelic lenited mh
ph	0.61	0.70	0.19	0.33	Gaelic lenited ph

first 30 ranks are extracted, which is sufficient to illustrate the qualitative contrast referenced throughout this volume's Chapter 11.

The contrast between the two tables is the bigram-level signature of the entropy and KL results reported in Chapter 11: the Anatolian top-30 frequencies decay smoothly from 2.62% to 1.40% (a ratio of 1.87 between rank 1 and rank 30), consistent with its comparatively high, evenly-spread bigram entropy, whereas the Armenian list is dominated by a single bigram (ar at 5.58%, more than double its Anatolian counterpart) before falling off more steeply to 1.01% by rank 30 (a ratio of 5.53). This concentration at the top of the Armenian distribution is precisely what depresses Armenian bigram entropy to the lowest value in the volume (7.540 bits, Table 11.1) and what generates its anomalously high self-perplexity-relative-to-others in Table 11.3.

Remark: Scope of This Appendix

A complete top-30 (or top-50) bigram table for every one of the 33 substrate zones catalogued in Table 11.1 would require the equivalent appendix from Vols. II, III, and V, none of which were supplied alongside Vols. IV, VI, and VII at the time this appendix was compiled. The tables above and below should be read as representative, fully data-backed illustrations of the method, not as exhaustive coverage of the series.

Table C.4: Top-30 bigrams, Anatolian interior zone ($n_{\text{bigrams}} \approx 25{,}500$; from Vol. VI, Table 9.1 / `tab:bigrams_anatolian`).

Rk	Bigram	%	Rk	Bigram	%
1	ar	2.62	2	si	2.60
3	li	2.32	4	ue	2.14
5	ka	2.10	6	al	2.08
7	la	2.04	8	er	2.02
9	ay	1.98	10	ak	1.94
11	an	1.92	12	ko	1.89
13	oy	1.87	14	de	1.85
15	da	1.83	16	te	1.81
17	re	1.78	18	as	1.74
19	in	1.70	20	ir	1.68
21	el	1.62	22	uk	1.59
23	ik	1.57	24	ta	1.54
25	ni	1.51	26	ya	1.49
27	ek	1.46	28	ke	1.44
29	ra	1.42	30	ba	1.40

C.5 Top-10 Bigrams by Zone: Maghreb (Vol. VII)

Table C.6 reproduces the top-10 bigram frequency table from Vol. VII (Table `tab:topbigrams` in that volume), covering all five Maghrebi zones. Values are percentage frequencies over the $n = 3{,}000$ zone corpus, seed 42.

The most diagnostically significant contrast in the Maghrebi top-10 tables is Libya's `qa` (3.09%), which is entirely absent from the top-10 of all other zones. The uvular stop q is phonemically preserved in Eastern Arabic (Libyan, Egyptian, Gulf varieties) but merged with $\varnothing$ or k in Western Maghrebi Arabic and in Berber-influenced zones. Morocco's dominance of `ou` (7.58%) reflects the Berber *ou-* prefix combined with the French transcription convention for the Berber high back vowel /u/. The tight DZ–TN cluster is already visible in the top-10 lists: both zones share `el`, `ou`, `ar`, `be`, `eb`, `je` in their top-10, with only minor reordering.

Table C.5: Top-30 bigrams, Armenian zone (from Vol. VI, `tab:bigrams_armenian`).

Rk	**Bigram**	**%**	**Rk**	**Bigram**	**%**
1	ar	5.58	2	ka	2.83
3	an	2.61	4	ha	2.54
5	sa	2.24	6	va	2.18
7	ra	2.12	8	na	2.07
9	ak	2.01	10	ts	1.95
11	av	1.89	12	al	1.83
13	nt	1.78	14	in	1.72
15	ta	1.67	16	at	1.62
17	ov	1.56	18	sh	1.51
19	kh	1.47	20	ut	1.42
21	gh	1.38	22	vn	1.33
23	ni	1.29	24	ot	1.25
25	it	1.21	26	er	1.16
27	ma	1.12	28	ne	1.08
29	li	1.05	30	ag	1.01

Table C.6: Top-10 bigrams by frequency (% of all bigrams in zone corpus) for the five Maghrebi zones (Vol. VII, real GeoNames corpus).

MA		DZ		TN		LY		MR	
Bigram	%	Bigram	%	Bigram	%	Bigram	%	Bigram	%
ou	7.58	ou	5.46	el	5.01	ra	5.69	el	4.49
ar	3.90	el	4.21	ou	4.84	al	5.24	ou	4.10
la	2.38	ar	2.37	be	2.96	ar	4.57	he	2.28
ua	2.27	be	2.37	ar	2.34	at	4.40	ei	2.12
do	1.96	dj	2.21	eb	2.23	ah	3.48	ha	1.81
ha	1.95	eb	1.82	je	2.08	ta	3.28	le	1.66
ra	1.89	je	1.75	ra	1.99	qa	3.09	ar	1.64
el	1.61	ha	1.64	ha	1.71	ha	3.00	eh	1.56
ta	1.60	ra	1.62	la	1.68	as	2.38	ra	1.48
ad	1.49	la	1.55	te	1.52	la	2.16	la	1.45

Asymptotic Theory of the Permutation Framework

This appendix develops the asymptotic distribution theory underlying the permutation tests used throughout the series (Section 3.3). It shows that, under mild regularity conditions, the permutation test for KL divergence is asymptotically equivalent to a likelihood ratio test, and derives the rate of convergence of the permutation distribution to its large-sample limit.

C.6 Setup and Notation

Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two toponymic corpora of sizes N_1 and N_2 respectively, drawn independently from distributions $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ over Σ^n. The null hypothesis of the permutation test is $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}$ for some common (unknown) distribution $\boldsymbol{\theta}$.

Define the *observed divergence statistic*

$$T_N \;=\; N_1 \cdot D_{\mathrm{KL}}(\hat{p}_1 \| \hat{p}_{12}) + N_2 \cdot D_{\mathrm{KL}}(\hat{p}_2 \| \hat{p}_{12}),$$

where $N = N_1 + N_2$, $\hat{p}_1$ and $\hat{p}_2$ are the empirical n-gram distributions of the two corpora, and $\hat{p}_{12}$ is the pooled empirical distribution with weight N_1/N on $\hat{p}_1$ and N_2/N on $\hat{p}_2$. Note that $T_N = N \cdot \mathrm{JS}_{N_1/N}(\hat{p}_1, \hat{p}_2)$ where JS_λ is the weighted Jensen–Shannon divergence with mixing weight λ.

C.7 Large-Sample Distribution Under the Null

C.7.1 Asymptotic Chi-Squared Limit

Theorem: Asymptotic Null Distribution of T_N

Assume $N_1/N \to \lambda \in (0,1)$ as $N \to \infty$, and that the common distribution $\boldsymbol{\theta}$ has full support over Σ^n (i.e. $\theta_g > 0$ for all g). Under H_0,

$$T_N \;\xrightarrow{d}\; \frac{1}{2}\chi^2(K-1) \quad \text{as } N \to \infty,$$

where $K = |\Sigma^n|$ is the number of n-gram types and $\chi^2(K-1)$ denotes the chi-squared distribution with $K-1$ degrees of freedom.

Proof sketch. Under H_0, the pooled count vector $\mathbf{c}_{12} = \mathbf{c}_1 + \mathbf{c}_2$ follows Mult$(N, \boldsymbol{\theta})$. Conditional on $\mathbf{c}_{12}$, the count vector $\mathbf{c}_1$ follows a multivariate Hypergeometric distribution Hyper$(N_1; \mathbf{c}_{12})$.

Expanding the KL divergences to second order in $\delta_1 = \hat{p}_1 - \boldsymbol{\theta}$ and $\delta_2 = \hat{p}_2 - \boldsymbol{\theta}$,

$$D_{\mathrm{KL}}(\hat{p}_1 \| \hat{p}_{12}) \;\approx\; \frac{1}{2} \sum_g \frac{(\hat{p}_1(g) - \hat{p}_{12}(g))^2}{\hat{p}_{12}(g)},$$

and similarly for the second term. The sum T_N becomes approximately

$$T_N \;\approx\; \frac{N}{2} \sum_g \frac{(\hat{p}_1(g) - \hat{p}_2(g))^2}{\hat{p}_{12}(g)} \cdot \lambda(1-\lambda),$$

which is proportional to the chi-squared statistic for the two-sample goodness-of-fit test with $K-1$ degrees of freedom (one constraint from the probability sum). The factor of $1/2$ arises because T_N is a weighted average of the two directional KL terms. By the multivariate central limit theorem applied to $\sqrt{N_1}(\hat{p}_1 - \boldsymbol{\theta})$ and $\sqrt{N_2}(\hat{p}_2 - \boldsymbol{\theta})$, the quadratic form converges in distribution to $\frac{1}{2}\chi^2(K-1)$ under the null, by the standard argument for Pearson's chi-squared statistic (with the pool providing the null estimate $\hat{p}_{12}$ of $\boldsymbol{\theta}$). □

Corollary: Consistency of the Permutation Test

The permutation test that rejects H_0 when T_N exceeds the $(1-\alpha)$-quantile of its permutation distribution is *consistent* against any fixed alternative $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$: i.e. the rejection probability converges to 1 as $N \to \infty$.

Proof. Under any fixed alternative with $D_{\mathrm{KL}}(\boldsymbol{\theta}_1 \| \boldsymbol{\theta}_2) > 0$, the observed statistic T_N grows at rate $O(N)$ (the sum of terms of order $N_1 \cdot D_{\mathrm{KL}}(\hat{p}_1 \| \hat{p}_{12})$ and $N_2 \cdot D_{\mathrm{KL}}(\hat{p}_2 \| \hat{p}_{12})$, each of which concentrates around positive constants by the law of large numbers). The permutation distribution, by contrast, concentrates around its null limit $\frac{1}{2}\chi^2(K-1)$ (by the permutation CLT, since under permutation the label assignment is exchangeable). Hence the ratio of the observed statistic to the critical value diverges to infinity, ensuring rejection with probability approaching 1. □

C.8 Rate of Convergence of the Permutation Distribution

C.8.1 Berry–Esseen Bound

Proposition: Berry–Esseen Rate for the Permutation Distribution

Let B_N denote the number of permutation replicates for which $T^{(b)} \geq T_N^{\mathrm{obs}}$. Under H_0 and the conditions of Theorem C.7.1, the Kolmogorov–Smirnov distance between the permutation distribution of T_N and its $\frac{1}{2}\chi^2(K-1)$ limit satisfies

$$\sup_{t \geq 0} \left| P_\sigma(T_N^{(\sigma)} \leq t) - F_{\frac{1}{2}\chi^2_{K-1}}(t) \right| \;=\; O\!\left(\frac{K^{3/2}}{N^{1/2}} \right),$$

where P_σ denotes the permutation distribution (over uniform random permutations σ of the N toponyms), and $F_{\frac{1}{2}\chi^2_{K-1}}$ is the CDF of the limiting distribution.

This rate implies that, at the bigram level ($K = 729$) and corpus sizes $N = 6{,}000$ (two zones of 3,000 each), the Kolmogorov–Smirnov distance is of order $729^{3/2}/6000^{1/2} \approx 19{,}683/77.5 \approx 254$, which sounds large in absolute terms but is effectively irrelevant in practice because the observed statistics under the series' alternatives are so far into the tail of the null distribution that the precise shape of the null tail does not affect the p-value. The approximation matters most when the observed T_N is near the critical quantile — a situation that does not arise in any of the within-volume comparisons of this series, where all pairwise p-values are either < 0.001 or > 0.30.

C.9 Multiple Testing Correction in the Context of Permutation Tests

When m pairwise permutation tests are conducted simultaneously (one for each pair of zones in a k-zone analysis, giving $\binom{k}{2}$ tests), the overall family-wise error rate (FWER) inflates with m. The present series uses the Benjamini–Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001) for FDR control, which is valid for arbitrarily dependent test statistics (as the permutation test statistics are, since they share the same pooled corpus).

Proposition: FDR Control Under the BY Procedure

For m simultaneously conducted permutation tests at nominal level α, the Benjamini–Yekutieli procedure controls the FDR at $\alpha \cdot \frac{1}{\sum_{i=1}^{m} 1/i}$ under any joint dependence structure of the test statistics. For $m = \binom{k}{2}$ with $k \leq 6$ (the maximum in this series, giving $m = 15$), the correction factor is $1/\sum_{i=1}^{15}(1/i) \approx 1/3.38 \approx 0.30$, so the effective FDR level achieved at nominal $\alpha = 0.05$ is ≈ 0.015.

This strong control is an appropriate choice in the series because the costs of a false discovery (incorrectly concluding that two zones have distinct phonotactic distributions when they do not) are asymmetric with the costs of a false non-discovery. A false discovery leads to the erroneous invocation of a substrate contrast that does not exist in the data; a false non-discovery leads only to the conservative conclusion that two zones are undistinguished — the safer error in the context of a quantitative framework that explicitly positions itself as a supplement to, rather than a replacement for, the etymological and historical-linguistic evidence.

C.10 Power Under Sparse Alternatives

C.10.1 Formulation of the Sparse Alternative

The foregoing power analysis (Section 3.4) considers the "dense" alternative in which all K bigram probabilities differ between the two zones. In the context of sub-

strate detection, however, the alternative is typically *sparse*: a substrate differs from the dominant language in a small number of characteristic n-grams (the marker bigrams of Appendix D), while the remaining n-grams have very similar probabilities. Sparse alternatives are harder to detect with aggregate statistics like D_{KL} or W_1 and require either larger corpus sizes or more sensitive test statistics.

Proposition: Power Under Sparse Alternatives

Suppose that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ differ on a set $\mathcal{S} \subset \Sigma^n$ of $s \ll K$ bigrams, with $|\theta_{1,g} - \theta_{2,g}| = \delta_0$ for $g \in \mathcal{S}$ and $\theta_{1,g} = \theta_{2,g}$ for $g \notin \mathcal{S}$. The expected KL divergence under this alternative is

$$\mathbb{E}\big[D_{\mathrm{KL}}(\hat{p}_1 \| \hat{p}_2)\big] \;\approx\; s \cdot \frac{\delta_0^2}{2\theta_0} + O\left(\frac{K + s^2\delta_0^2/\theta_0}{N}\right),$$

where θ_0 is the common baseline probability of the s affected bigrams. The permutation test achieves power $1 - \beta$ at sample size N when

$$N \;\gtrsim\; \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2\, \theta_0}{s\, \delta_0^2},$$

i.e. when N is inversely proportional to $s\delta_0^2/\theta_0$, the *total signal-to-noise* of the sparse alternative.

For the marker bigram contrasts of Appendix D, $s \approx 5$–10 marker bigrams per zone pair, $\delta_0 \approx 0.01$ (1 percentage point shift), and $\theta_0 \approx 0.01$–0.02. Substituting into the formula gives $N \gtrsim 2(1.96 + 0.84)^2 \times 0.015/(5 \times 10^{-4}) = 2 \times 7.84 \times 0.015/5 \times 10^{-4} \approx 470$ per zone — well below the corpus sizes of 3,000 per zone used in this series, confirming adequate power for sparse marker-bigram contrasts. For trigram alternatives ($n = 3$) where θ_0 is typically an order of magnitude smaller ($\approx 10^{-3}$–10^{-4}), the required sample size increases by the same factor, reaching $4{,}700$–$47{,}000$ per zone for the same signal strength, which explains the observed loss of sensitivity in the trigram analyses of the regional volumes relative to the bigram core.

C.11 Jackknife Variance Estimation for the KL Divergence

C.11.1 Definition and Computational Cost

An alternative to the bootstrap for estimating the sampling variance of $\hat{D}_{\mathrm{KL}} = D_{\mathrm{KL}}(\hat{p}_1 \| \hat{p}_2)$ is the *delete-one jackknife*, which has the advantage of being analytically tractable for smooth statistics and is faster to compute at large N.

Definition: Delete-One Jackknife for KL Divergence

For a corpus $\mathcal{T}$ of size N, the *delete-one jackknife replicate* is the statistic computed after removing the i-th toponym w_i:

$$\hat{D}^{(-i)} \;=\; D_{\mathrm{KL}}(\hat{p}_1^{(-i)} \| \hat{p}_2),$$

where $\hat{p}_1^{(-i)}$ is the empirical distribution of $\mathcal{T}_1 \setminus \{w_i\}$. The jackknife variance estimate is

$$\widehat{\mathrm{Var}}_{\mathrm{JK}}(\hat{D}) \;=\; \frac{N-1}{N}\sum_{i=1}^{N}\big(\hat{D}^{(-i)} - \bar{D}^{(\cdot)}\big)^2, \quad \bar{D}^{(\cdot)} = \frac{1}{N}\sum_{i}\hat{D}^{(-i)}.$$

Proposition: Consistency of the Jackknife Variance

Under the Multinomial model for n-gram counts, the jackknife variance estimator $\widehat{\mathrm{Var}}_{\mathrm{JK}}(\hat{D})$ is a consistent estimator of the true variance of $\hat{D}_{\mathrm{KL}}$ as $N \to \infty$, at the same rate as the bootstrap estimator ($O_p(N^{-1})$), but without requiring the specification of the number of bootstrap replicates B. Its computational cost is $O(NK)$ elementary operations (one partial n-gram count update per toponym removal), compared to $O(BNK)$ for the bootstrap with B replicates.

In the present series, the bootstrap ($B = 1{,}000$) is retained as the primary variance-estimation procedure throughout because it is more familiar to the intended readership and its properties are well-documented in the specific context of entropy estimation (Efron, 1987). The jackknife is available as a computationally efficient alternative and is used in the robustness analyses of Chapter 9, where variance estimation must be repeated across hundreds of parameter configurations.

C.11.2 Comparison with Bootstrap Variance at Finite N

For corpus sizes $N \geq 3{,}000$ and the bigram alphabet ($K = 729$), both the bootstrap and the jackknife produce variance estimates that are within 5% of each other in all pairwise comparisons conducted in this series. The jackknife is slightly anti-conservative (tends to underestimate variance) near the zero boundary of D_{KL} — the same boundary at which the bootstrap is biased for the opposite reason (truncation at zero inflates the bootstrap quantile). For the predominant case in this series, where $D_{\mathrm{KL}} \gg 0.1$ bits, both estimators are effectively unbiased and the choice between them is a matter of computational convenience.

C.12 Summary and Practical Recommendations

The results of this appendix can be summarised in four practical recommendations for the permutation testing pipeline:

(i) **Use $B \geq 1{,}000$ permutation replicates.** At $B = 1{,}000$, the Monte Carlo standard error of the estimated p-value at $p = 0.05$ is $\sqrt{0.05 \times 0.95/1000} \approx 0.007$, small relative to any substantive threshold.

(ii) **Apply the BY correction for all within-volume pairwise tests.** The effective FDR level of ≈ 0.015 at nominal $\alpha = 0.05$ is appropriate for the series' inferential goals.

(iii) **Report bootstrap CIs alongside permutation p-values.** The asymptotic theory shows that the two procedures are complementary: the permutation test controls type I error at finite N; the bootstrap CI quantifies estimation uncertainty around the effect size.

(iv) **For trigram analyses, use the NSB entropy estimator** (Section 3.2.2) rather than Miller–Madow, since the bigram-regime approximation $N/K \gg 1$ breaks down for trigrams at the corpus sizes of this series.

Bibliography

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the 18th Annual ACM-SIAM SODA*, pp. 1027–1035.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4), 1165–1188.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Concepts and Techniques in Modern Geography, No. 38. Geo Books, Norwich.

Bentz, C., Alikaniotis, D., Cysouw, M., and Ferrer-i-Cancho, R. (2017). The entropy of words — learnability and expressivity across more than 1000 languages. *Entropy*, 19(6):275.

Cliff, A.D. and Ord, J.K. (1981). *Spatial Processes: Models and Applications*. Pion, London.

Cressie, N. (1993). *Statistics for Spatial Data*, revised edition. Wiley, New York.

Moran, P.A.P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Cavnar, W.B. and Trenkle, J.M. (1994). N-gram-based text categorization. *Proceedings of SDAIR-94*, 3:161–175.

Cysouw, M. and Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *STUF*, 60(2):95–99.

Dryer, M.S. and Haspelmath, M. (eds.) (2013). *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dunning, T. (1994). Statistical identification of language. Technical Report MCCS-94-273, Computing Research Laboratory, New Mexico State University.

Endres, D.M. and Schindelin, J.E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860.

Grassberger, P. (1988). Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6–7):369–373.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Krahe, H. (1964). *Unsere ältesten Flussnamen*. Harrassowitz, Wiesbaden.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv*: 1802.03426.

Miller, G.A. (1955). Note on the bias of information estimates. In *Information Theory in Psychology*, pp. 95–100. Free Press.

Monaghan, P., Shillcock, R.C., Christiansen, M.H., and Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B*, 369(1651):20130299.

Nemenman, I., Shafee, F., and Bialek, W. (2002). Entropy and inference, revisited. *Advances in Neural Information Processing Systems*, 14:471–478.

Nicolaisen, W.F.H. (1976). *Scottish Place-Names: Their Study and Significance*. B.T. Batsford, London.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253.

Pellegrini, G.B. (1990). *Toponomastica italiana*. Hoepli, Milan.

Pulgram, E. (1954). Theory of names. *Beiträge zur Namenforschung*, 5:149–196.

Salvador, S. (2024). Lambioi: note di toponomastica bellunese medievale. *Archivio Storico Bellunese*.

Salvador, S. (2024a). *Toponomastic Entropy, Vol. I: Methodology and Information-Theoretic Framework*. Independent monograph series.

Salvador, S. (2024b). *Toponomastic Entropy, Vol. II: Iberian Peninsula*. Independent monograph series.

Salvador, S. (2024c). *Toponomastic Entropy, Vol. III: Central Europe*. Independent monograph series.

Salvador, S. (2024d). *Toponomastic Entropy, Vol. IV: British Isles*. Independent monograph series.

Salvador, S. (2024e). *Toponomastic Entropy, Vol. V: Indian Subcontinent*. Independent monograph series.

Salvador, S. (2024f). *Toponomastic Entropy, Vol. VI: Anatolia and the Caucasus*. Independent monograph series.

Salvador, S. (2025). *Toponomastic Entropy, Vol. VII: Maghreb*. Independent monograph series.

Shannon, C.E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64.

Wichmann, S., Holman, E.W., Bakker, D., and Brown, C.H. (2010). Evaluating linguistic distance measures. *Physica A*, 389(17):3632–3639.

Rényi, A. (1961). On measures of entropy and information. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, 1:547–561.

van Erven, T. and Harremoës, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Villani, C. (2009). *Optimal Transport: Old and New*. Springer, Berlin.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300.